
WASSERSTEIN GRADIENT FLOW OF TWO-LAYERS ReLU NETWORKS

Victor Mercklé
August 27, 2024

CONTENTS

| | |
|--|-----------|
| 1 Two-layers ReLU | 3 |
| 1.1 The model | 3 |
| 1.2 Training | 3 |
| 1.3 Explicit gradient computations | 3 |
| 2 Convergence of mean-field networks | 4 |
| 2.1 Algorithm, discretize the measure's space | 5 |
| 2.2 Mean field rewritten results | 5 |
| 2.2.1 direct result | 5 |
| 2.2.2 simple result | 5 |
| 3 Simple example 2D setting, grid jko | 5 |
| 4 Proximal point - direct WasserImplemented descents | 7 |
| 4.1 Gradient descent | 7 |
| 4.2 Proximal point | 7 |
| 4.2.1 dist = L2 sq, frobenius | 7 |
| 4.2.2 dist = wasserstein2 sq | 8 |
| 5 Prox point: Comparing the objective value and distance between iterates | 9 |
| 5.1 Early experiments | 9 |
| 5.1.1 proximal point | 9 |
| 5.1.2 wasserstein descent | 10 |
| 6 Larger simulations on Prox point | 11 |
| 7 What's the difference between wasser and frobenius prox: 2 pts example | 13 |
| 7.1 détail wasser | 14 |
| 7.2 wasserstein: détails de la première étape | 14 |
| 7.3 Extensions | 16 |
| 7.4 Next steps | 17 |
| 8 intuition on Small vs Big initialization in ReLU | 17 |

| | |
|---|-----------|
| 9 JKO | 17 |
| 9.1 JKO stepping with Dykstra's algorithm | 17 |
| 9.2 Papers | 18 |

1 TWO-LAYERS RELU

1.1 THE MODEL

Data $x_j \in \mathbb{R}^d$ and labels $y_j \in \mathbb{R}$, $j = 1, \dots, n$ (The network has a scalar output).

First ReLU layer parameters: $w_i \in \mathbb{R}^d$, second linear layer parameters: $\alpha_i \in \mathbb{R}$ for $i = 1, \dots, m$. The bias is implemented by adding a dimension to the data full of ones, which is strictly equivalent for a two-layers network.

$\beta > 0$ regularization constant.

$$\mathcal{L}(W, \alpha) = \underbrace{\sum_{j=1}^n \left(\sum_{i=1}^m \max(0, w_i^\top x_j) \alpha_i - y_j \right)^2}_{\text{Network's Output}} + \underbrace{\beta \sum_{i=1}^m \|w_i\|_2^2 + \alpha_i^2}_{\text{Weight Decay}} \quad (1)$$

This is an explicit form of using the mean squared error (MSE) loss along with L_2 regularization.

1.2 TRAINING

Discret time.

Full-batch gradient descent with $\gamma > 0$ the step-size (or learning rate).

$$(W, \alpha)_{t+1} = (W, \alpha)_t - \gamma \nabla \mathcal{L}((W, \alpha)_t) \quad (2)$$

Implicit ($\theta = (W, \alpha) \in (\mathbb{R}^{m \times d}, \mathbb{R}^m)$)

$$\theta_{t+1} = \arg \min_{\theta} \mathcal{L}(\theta) + \frac{1}{2\gamma} \|\theta - \theta_t\|_2^2 \quad (3)$$

Continuous time.

Taking $\gamma \rightarrow 0$, we get the gradient flow: $\frac{d\theta_t}{dt} = -\nabla \mathcal{L}(\theta_t)$. We can make ReLU differentiable with $\sigma'(0) = 0$ as justified in (Boursier et al., 2022).

Given a smooth function $a \rightarrow F(a)$, the gradient flow is gradient descent algorithm

$$a^{l+1} = a^l - \gamma \nabla F(a^l)$$

with a small enough γ . If F is not smooth, the gradient flow is the proximal-point algorithm

$$a^{l+1} = \text{Prox}_{\gamma F}^{\|\cdot\|}(a^l) = \arg \min_a \frac{1}{2} \|a - a^l\|^2 + \gamma F(a)$$

with a small enough γ .

If F is defined on histograms, it makes sense to use the wasserstein distance W^p

1.3 EXPLICIT GRADIENT COMPUTATIONS

Here we will fix the second layer to be in $-1, 1$ as is done in mean-field and JKO and to focus on the most important part of the dynamic. Explicit gradient of the second layer is given later.

Take a two layer ReLU network with m neurons. Each neuron has a trainable parameter $w_i \in \mathbb{R}^d$ and a fixed output sign $\alpha_i \in \{-1, 1\}$. Each of the n data points of dimension $d - 1$ are augmented with a 1, so each sample is of dimension n . Each sample x_j is stored as a row $X \in \mathbb{R}^{n \times d}$ and is associated with a scalar label $y_j \in \mathbb{R}$.

The output of one neuron is: $x \in \mathbb{R}^d \rightarrow \max(0, \langle w_i, x \rangle) \alpha_i$, shorthand $\langle w_i, x \rangle_+ \alpha_i$.

The output of a network of m neurons on one data point is $f(x) = \sum_{i=1}^m \max(0, \langle w_i, x \rangle) \alpha_i$

We store the neuron trainable parameter w_i as the columns of $W \in \mathbb{R}^{d \times m}$. The loss for n data points

$$F(W) = \frac{1}{n} \sum_{j=1}^n (f(x_j) - y_j)^2 \quad (4)$$

Discretized, unregularized gradient descent with $\lambda \in \mathbb{R}^+$ stepsize: (resolving the non differentiable points $\max(0, 0)$ with 0 as gradient)

$$W^{t+1} = W^t - \lambda \nabla F(W^t) \quad (5)$$

Taking $\lambda \rightarrow 0$, we get the gradient flow.

Explicit discret gradient:

$$\frac{\partial F}{\partial w_i^t} = \frac{\alpha_i^t}{n} \sum_{j=1}^n e_j^t s_{i,j}^t x_j \quad (6)$$

With real $e_j^t = f(W^t) - y_j$ the "signed error on input j "

With boolean $s_{i,j}^t = \mathbb{1}_{\langle w_i^t, x_j \rangle > 0}$ the "is neuron i activating on datapoint j ". The vector $s_i^t \in \{0, 1\}^n$ would be the activation pattern of neuron i at time t .

Remark: the gradient of a neuron is a linear combination of the data points it activates.

Below is the explicit gradient for the second layer (if we trained it as well).

$$\frac{\partial F}{\partial \alpha_i^t} = \frac{1}{n} \sum_{j=1}^n e_j^t s_{i,j}^t \langle w_i^t, x_j \rangle \quad (7)$$

2 CONVERGENCE OF MEAN-FIELD NETWORKS

Mean-field limit(Chizat & Bach): For a sufficiently large width, the training dynamics of a NN can be coupled with the evolution of a probability distribution described by a PDE.

If [...] converges, with $m \rightarrow \infty$ (many-particle limit), our particles of interest converges to a Wasserstein gradient flow of F:

$$\partial \mu_t = -\text{div}(v_t \mu_t) \text{ where } v_t \in -\partial F'(\mu_t) \quad (8)$$

$$\int_{\Theta} m(\theta; x) d\mu(\theta) = \frac{1}{m} \sum_{i=1}^m \langle w_i, x \rangle_+ \alpha_i \quad (9)$$

Remark: le $1/m$ c'est principalement pour être ok à l'infini. Dans le papier JKO (Carlier et al.) ils utilisent un vecteur de proba

Different ways to use a measure to represent the neurons of a two layer network:

- $\Theta = \mathbb{R}^d \times \mathbb{R}$, measure $\mu = \frac{1}{m} \sum_{i=1}^m \delta_{\theta_i=(w_i, \alpha_i)}$, output of one neuron $m(\theta = (w, \alpha); x) = \langle x, w \rangle_+ \alpha$: (works, output matches discrete)

- $\Theta = \mathbb{R}^d$, measure $\mu = \frac{1}{m} \sum_{i=1}^m \alpha_i \delta_{\theta_i = w_i}$ output of one neuron $m(\theta = w; \mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle_+$ (works)
- $\Theta = \mathbb{R}^d \times \mathbb{R}^d$, output of one neuron $m(\tilde{\mathbf{w}}_+, \tilde{\mathbf{w}}_-, \mathbf{x}) = \langle \tilde{\mathbf{w}}_+, \mathbf{x} \rangle - \langle \tilde{\mathbf{w}}_-, \mathbf{x} \rangle$ (works, separate neg and positive)
- $\Theta = (S^{d-1} \times \mathbb{R})$, output of one neuron $m((\mathbf{d}, \tilde{\alpha}); \mathbf{x}) = \tilde{\alpha} \langle \mathbf{d}, \mathbf{x} \rangle = \tilde{\alpha} \mathbb{1}_{\langle \mathbf{d}, \mathbf{x} \rangle > 0}$ (works), mapping: $\mathbf{d} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$ and $\tilde{\alpha} = \|\mathbf{w}\| \alpha$. Gradient are not equal to discrete.

2.1 ALGORITHM, DISCRETIZE THE MEASURE'S SPACE

Take a grid of N points in Θ , we can match the notation above by taking a neuron for each point of the grid $m = N$.

The grid currently dictate the neuron's scale, giving multiple choices.

$$\mu(t+1) = \arg \min_{\mu \in \mathcal{M}(\Theta)} F(\mu) + \frac{1}{2\gamma} W_2(\mu; \mu(t)) \quad (10)$$

Take $\gamma \rightarrow 0$, get gradient flow. Take $m \rightarrow \infty$, get wasserstein gradient flow (Bach & Chizat), and if it converges, it goes to the global optimal.

2.2 MEAN FIELD REWRITTEN RESULTS

2.2.1 DIRECT RESULT

Minimization of a linear combination of ('neurons' $\{\phi(\theta)\}_{\theta \in \Theta}$) through an unknown measure μ : $J^* = \min_{\mu \in \mathcal{M}(\Theta)} J(\mu) = R(\int \phi d\mu) + G(\mu)$

With R a convex loss function and G a convex regularizer, $\mathcal{M}(\Theta)$ the set of signed measures on the parameter space Θ .

Discretize the measure into m particles: $\min_{w \in \mathbb{R}^m} \theta \in \Theta^m J_m(w, \theta) = J(\frac{1}{m} \sum_{i=1}^m w_i \delta_{\theta_i})$

Proved: if WGF cvg, it cvg to global minimizer. If $(w^{(m)}(t), \theta^{(m)}(t))_{t \geq 0}$ are gradient flows for J_m then, with the corresponding measure $\mu_{m,t} = \frac{1}{m} \sum_{i=1}^m w_i^{(m)}(t) \sigma_{\theta_i^{(m)}(t)}$ (a WGF), $J(\mu_{m,t})$ cvg (with $m, t \rightarrow \infty$) to global minimizer of J .

p-Wasserstein distance between two measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$: $W_p(\mu, \nu)^p = \min_{\gamma \in \Pi(\mu, \nu)} \int |y - x|^p d\gamma(x, y)$

2.2.2 SIMPLE RESULT

Cast the parameters $w \in \mathcal{W}^{d+1}$ into $r \in \mathbb{R}, \eta \in S^d = w \in \mathbb{R}^{d+1}, \|w\|_2 = 1$ with $w = r \eta$. The two flows will have exactly the same dynamics.

Then measure $\nu = \frac{1}{m} \sum_{j=1}^m r_j^2 \sigma_{\eta_j}$ satisfy this PDE : $\partial_t \nu_t(\eta) = -4J(\eta|\nu_t) + \text{div}(\nu_t(\eta) \nabla J(\eta|\nu_t))$ (so it's a WGF)

Prediction functions : $h = \frac{1}{m} \sum_{j=1}^m \psi(w_j) = \frac{1}{m} \sum_{j=1}^m r_j^2 \psi(\eta_j)$

Theorem 1: if by taking $m \leftarrow \infty$ at $t = 0$, ν converges to ν_0 , then for any t , $\nu_{m,t}$ cvg to the unique WGF ν_t .

Theorem 2: Take a ν_0 with a support that includes all directions at initialization, if WGF ν_t cvg, it's to a global optimum.

3 SIMPLE EXAMPLE 2D SETTING, GRID JKO

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 $(x, 1) \in \mathbb{R}^2$:

$$w, b, \alpha \in \mathbb{R} \rightarrow \max(0, wx + b)\alpha \quad (11)$$

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

$$\mathcal{L} = \sum_{j=1}^n \left(\left(\sum_{i=1}^m \max(0, w_i x_j + b_i) \alpha_i \right) - y_j \right)^2 \quad (12)$$

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:

$$\int_{\mathbb{R}^2} m((w, b); x) d\mu((w, b)) \quad (13)$$

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 second layer (α):

$$\max(0, w_i x + b_i) \alpha_i = \max(0, |\alpha_i| (w_i x + b_i)) \text{sign}(\alpha_i) \quad (14)$$

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:

$$\int_{\mathbb{R}^2} m((w, b); x) d\mu((w, b)) = \sum_{i=1}^m \max(0, w_i x + b_i) \alpha_i p_i \quad (15)$$

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 movement 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:

$$\mu(t+1) = \arg \min_{\mu \in \mathcal{M}(\Theta)} F(\mu) + \frac{1}{2\gamma} W_2(\mu; \mu(t)) \quad (16)$$

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.

4 PROXIMAL POINT - DIRECT WASSERIMPLEMENTED DESCENTS

Discretized wasserstein prox step:

$$W^{t+1} = \arg \min_{W \in \mathbb{R}^{d \times m}} F(W) + \frac{1}{2\gamma} W_p(W; W^t) \quad (17)$$

EMD(POT library), d=squareeuclidian

$$W_2^2(W; W^t) = \min_{\gamma} \langle \gamma, M \rangle_F \quad (18)$$

$$\text{s.t. } \gamma \mathbf{1} = W \quad (19)$$

$$\gamma^\top \mathbf{1} = W^t \quad (20)$$

$$\gamma \geq 0 \quad (21)$$

$$M_{i,j} = \|w_i - w_j\|_2^2 \quad (22)$$

We only update the first layer $W^{(0)} \in \Theta = \mathbb{R}^{d \times m}$ for simplicity. $F : \Theta \rightarrow \mathbb{R}^+$ is the objective to minimize.

$$\min_{W \in \Theta} F(W) \quad (23)$$

4.1 GRADIENT DESCENT

$$W^{(t+1)} = W^{(t)} - \gamma \nabla F(W^{(t)}) \quad (24)$$

4.2 PROXIMAL POINT

$d : \Theta \times \Theta \rightarrow \mathbb{R}^+$ a distance between two parameters

$$W^{(t+1)} = \arg \min_{V \in \Theta} F(V) + \frac{1}{2\gamma} d(V, W^{(t)}) \quad (25)$$

To solve each arg min, we use a gradient descent on V . $V^{(0)} = W^{(t)}$

$$V^{(k+1)} = V^{(k)} - \beta \left(\nabla F(V^{(k)}) - \frac{1}{\gamma} \nabla d(V^{(k)}, W^{(t)}) \right) \quad (26)$$

4.2.1 DIST = L2 SQ, FROBENIUS

Frobenius norm ($L_{2,2}$): $\|W\|_{2,2}^2 = \|W\|_F^2 = \sum_{i=1}^m \sum_{j=1}^d W_{i,j}^2 = \sum_{i=1}^m \|W_i\|_2^2$

$$d : (V, W) \in \Theta \times \Theta \rightarrow \frac{1}{m} \sum_{i=1}^m \|V_i - W_i\|_2^2 \quad (27)$$

To sum up, we're solving this

$$W^{(t+1)} = \arg \min_{V \in \Theta} F(V) + \frac{1}{2\gamma} \frac{1}{m} \sum_{i=1}^m \|V_i - W_i\|_2^2 \quad (28)$$

Using gradient descent for arg min, one inner step (with inner step size β) on V with $V^{(0)} = W^{(t)}$ is

$$V^{(k+1)} = V^{(k)} - \beta \left(\nabla F(V^{(k)}) + \frac{1}{m \gamma} (V^{(k)} - W^{(t)}) \right) \quad (29)$$

$$V^{(k+1)} = V^{(k)} \left(1 - \frac{\beta}{m \gamma} \right) + \frac{\beta}{m \gamma} W^{(t)} - \beta \nabla F(V^{(k)}) \quad (30)$$

which is kinda the same thing as this (only for $\gamma \rightarrow 0$) (trust region problem equivalence with proximal problem) $V^{(0)} = W^{(t)}$

$$V^{(k+1)} = W^{(t)} - \gamma \nabla F(V^{(k)}) \quad (31)$$

4.2.2 DIST = WASSERSTEIN2 SQ

distance definition.

$$W^{(t+1)} = \arg \min_{V \in \Theta} F(V) + \frac{1}{2 \gamma} d(V, W^{(t)}) \quad (32)$$

With $V, W \in \mathbb{R}^d$ parameters.

p-Wasserstein distance between two measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$:

$$W_p(\mu, \nu)^p = \inf_{\gamma \in \Pi(\mu, \nu)} \int |y - x|^p d\gamma(x, y) \quad (33)$$

Π set of probability measures on \mathbb{R}^d with marginals μ and ν .

We take two sum of diracs measures $\mu = \frac{1}{m} \sum_{i=1}^m \delta_{V_i}$ and $\nu = \frac{1}{m} \sum_{i=1}^m \delta_{W_i}$. Then in that case W_2 is

$$W_2(V, W)^2 = \min_{\sigma \in S_n} \frac{1}{m} \sum_{i=1}^m \|V_i - W_{\sigma(i)}\|_2^2 \quad (34)$$

With σ a permutation. This is the linear assignment problem which can be solved by the hungarian algorithm. It always has a solution so inf=min

This is equivalent to what `OT.emd(d=squareeuclidian)` solves:

$$W_2^2(W; W^t) = \min_{\gamma} \langle \gamma, M \rangle_F \quad (35)$$

$$\text{s.t. } \gamma \mathbf{1} = W \quad (36)$$

$$\gamma^\top \mathbf{1} = W^t \quad (37)$$

$$\gamma \geq 0 \quad (38)$$

$$M_{i,j} = \|w_i - w_j\|_2^2 \quad (39)$$

However, we only need the gradient of the distance. I haven't yet read what it is exactly.

Small stepsize.

To sum up, for wasserstein descent we have this step

$$W^{(t+1)} = \arg \min_{V \in \Theta} F(V) + \frac{1}{2 \gamma m} \min_{\sigma \in S_n} \sum_{i=1}^m \|V_i - W_{\sigma(i)}\|_2^2 \quad (40)$$

If γ is small, the candidate V will be close to $W^{(t)}$, and each V_i will be close to $W_i^{(t)}$. Therefore the optimal assignment is the identity.

$$W^{(t+1)} = \arg \min_{V \in \Theta} F(V) + \frac{1}{2\gamma m} \sum_{i=1}^m \|V_i - W_i\|_2^2 \quad (41)$$

Which is exactly the proximal point. Additionnaly, as the step size gets smaller, proximal point converges to the gradient flow (just like gradient descent).

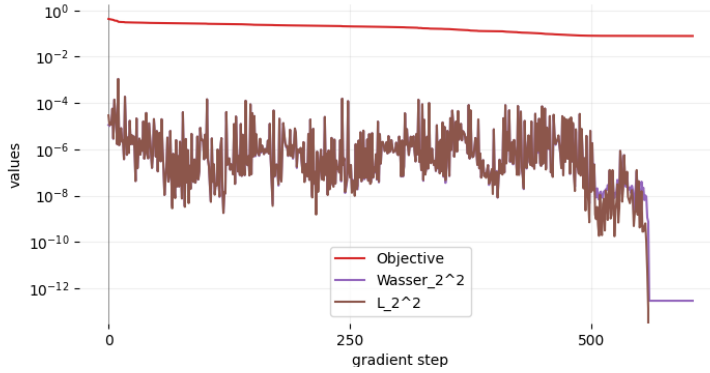
5 PROX POINT: COMPARING THE OBJECTIVE VALUE AND DISTANCE BETWEEN ITERATES

Which are the two items to minimize in proximal point.

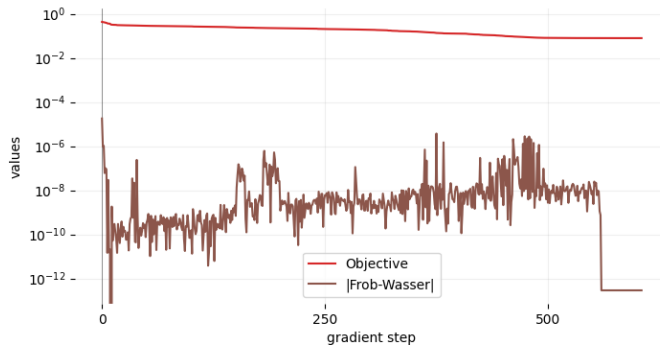
5.1 EARLY EXPERIMENTS

5.1.1 PROXIMAL POINT

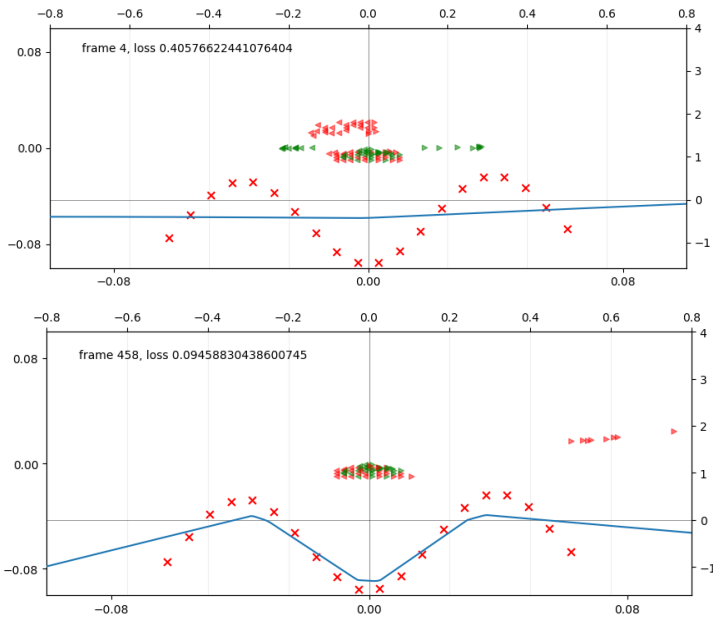
We run proximal point which optimize the objective $F(W^{t+1})$ and the distance between two iterates $d(W^{t+1}, W^{(t)})$. We plot those two values for $\gamma = 10$, and for Frobenius distance and Wasserstein distance as defined later.



It looks like they are exactly the same. We also plot the difference between the two distances to confirm.

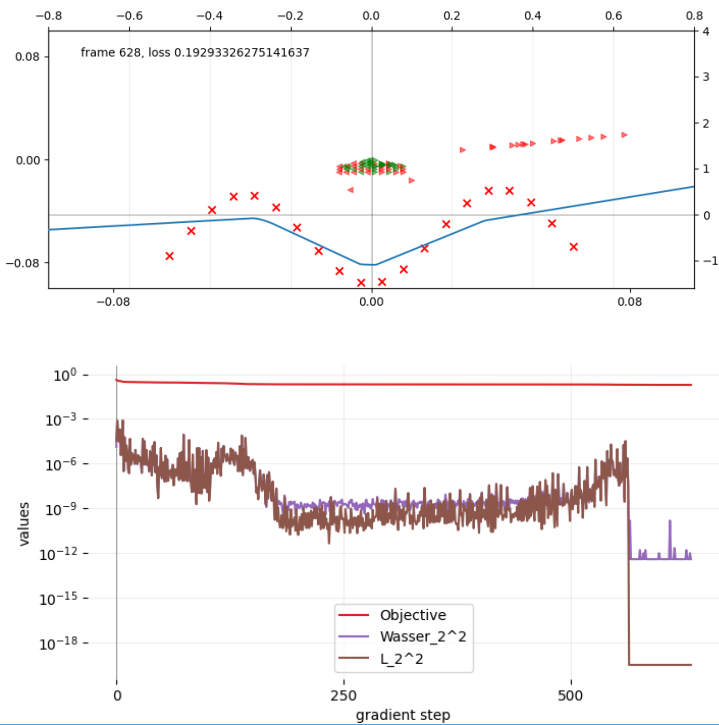


Start and 'end' of convergence for proximal point:



It's a relatively large stepsize but not an easy problem. The triangle of neurons that do not move are simply neurons that do not activate any data. Initialization is a random grid, with an output norm close to zero.

5.1.2 WASSERSTEIN DESCENT



The early trajectory is very similar, but then it looks stuck for a while.

Caveat: prox solution is not accurate due to small number of inner iterations.

6 LARGER SIMULATIONS ON PROX POINT

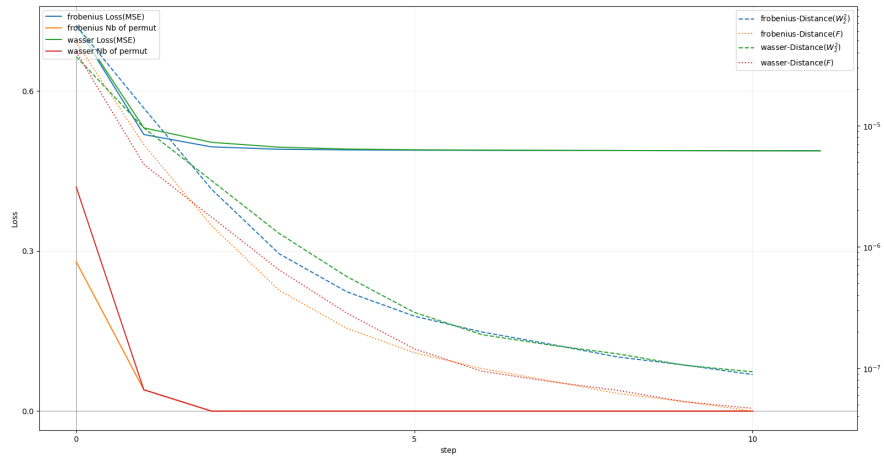
$m = 50$ (scale 10^{-2}), $n = 5$ (some usual $[-1, 1]$ sinus data), we compute 10^6 prox inner iterations (the bigger the stepsize the larger the step size might need to be) for both the frobenius and the wasserstein distance.

In blue/green filled lines we plot the normal MSE loss of the network (for frobenius proxpoint and wasserstein proxpoint)

In dotted/dashed lines we compute $d(W^t, W^{(t+1)})$ (distance between iterates, which is one of the two objectives minimized at each step) for d =wasserstein and d =frobenius.

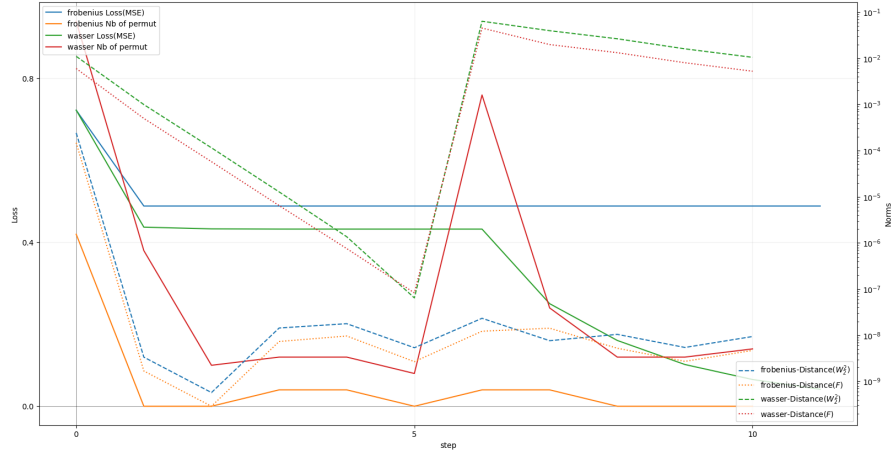
Because the only difference between wasserstein and frobenius is the assignment (which neuron/particle is assigned which particle to compute the distance), we also plot in thick green and red lines the number of permutations of the optimal mapping between two iterates. (1 = every particles has been permuted, 0 = each particle is associated to itself)

For $\gamma = 10^{-3}$ (small stepsize):



The 3 lines associated with wasserstein descent are very close to the 3 lines of frobenius descent, as expected.

For $\gamma = 10^0$ (big stepsize). This is one example where the convergence is very different. Wasserstein descent converges to 0 while Frobenius descent gets stuck on a bad local minima. We do the same plot as before.

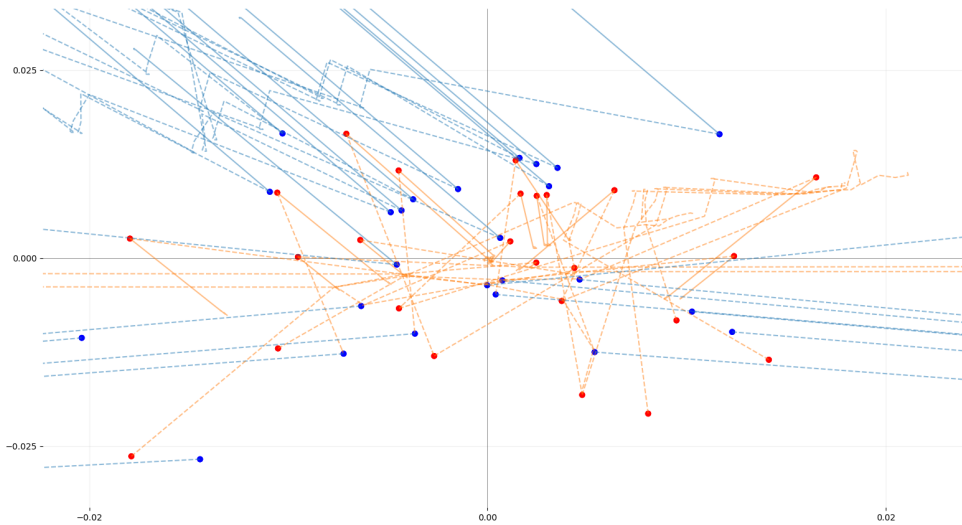


We can see the loss is completely different after 10 steps (it's normal to converge very quickly as the stepsize of the prox is large).

One thing to note, if we focus on wasserstein descent lines (dashed green and dotted red), the distance between iterates in frobenius or wasserstein is very close. However we can see in filled red that the number of permutations is close to 100% at some steps.

This indicates that the actual distance value is not as important as the actual trajectories of the neurones which are decided by the gradient of the wasserstein distance used in the prox.

We plot the trajectories of each neuron (dashed lines for wasserstein descent and filled lines for frobenius descent) for the 10 prox steps. Circles denotes the initial positions of the neurones, blue for negative neurones and red for positives neurones.

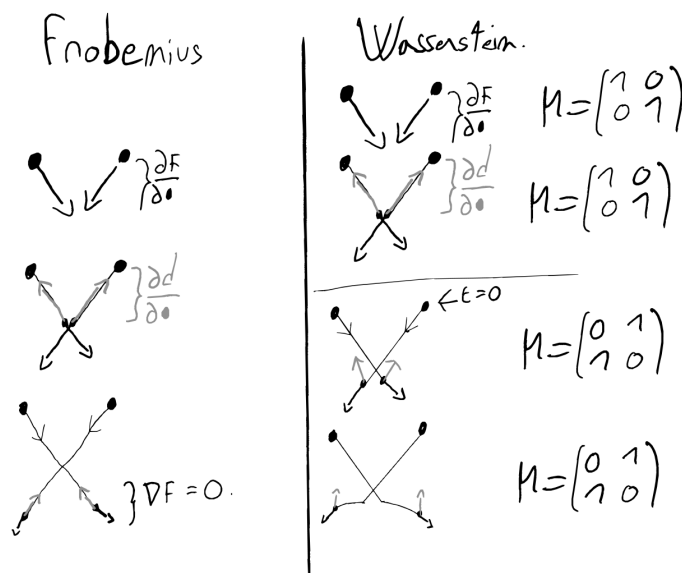


The trajectories are indeed different between the two descents. One major difference: neurons in the bottom quadrants are not moved at all by the frobenius descent, but in the wasserstein descent they are moved, they exchange positions (we can see two (dashed) trajectories very close going to and from the same initial neuron).

Those neurons (bottom/center quadrants) do not activate any datapoint and as such their differential is 0, so they cannot possibly move under prox descent. If weight decay is used, they will converge to 0. This completely differ from wasserstein descent, so the differential of the wasserstein distance must be different.

We'll focus on one proximal step by showing the first few *inner* iterations for the two descents. We take two particles with different gradients. (this could happen near a non-differentiability, or as we see in the simulations if we can switch positive and negative neurons)

The inner iterations are dictated by the gradient of the objective $\frac{\partial F}{\partial w_i}$ to minimize the loss, and by gradient of the distance to the first iterate ($W^{(0)}$). The frobenius proximal point goes as usual, the distance gradient pulling the particle back to its initial position. However the wasserstein distance ends up pulling the particles to the other neuron's initial positions.



This explains a scenario where the prox will converge to something different. However it doesn't explain directly why the wasserstein prox can move neurons that have 0 gradient.

Remark(not 100% sure): in the mean field paper, the only results on ReLU are obtained by computing the wasserstein distance on two separate sets (i.e. you cannot assign a positive neuron to a negative neuron)

Implementations details: Frobenius is ran on one GPU(15min per run) while wasserstein is ran on a single CPU core (30min per run) as that was found to be the most efficient to run multiple experiments. Results for $\gamma = 10^a$, $a = [0, -1, -2, -3, -4, -5]$ are available but do not differ much.

7 WHAT'S THE DIFFERENCE BETWEEN WASSER AND FROBENIUS PROX: 2 PTS

EXAMPLE

On étudie un problème en une dimension, avec analogie possible entre l'évolution de deux neurones ReLU(à activation fixe) dans \mathbb{R}^d et l'évolution de x (qui active une donnée et estime un label '4' et y qui n'active aucune donnée.

$$(x^{(0)}, y^{(0)}) = 0, 1$$

$$\min_{x,y \in \mathbb{R}} \underbrace{\frac{1}{2}(x-4)^2}_{f(x,y)} \quad (42)$$

En étudiant en détail une étape de prox, càd:

$$(x^{(t+1)}, y^{(t+1)}) = \arg \min_{(x,y)} f(x,y) + \frac{1}{2\gamma} d((x,y), (x^{(t)}, y^{(t)})) \quad (43)$$

Avec $\gamma > 0$ la taille du pas qui caractérise en gros à quel point la solution de l'argmin $((x,y)^{(t+1)})$ peut s'éloigner de l'initialisation/itéré précédent $((x,y)^{(t)})$

petit résumé de la suite.

On trouve que: avec $d = \text{frob/eucli}$ (i.e. $(x - x^{(t)})^2 + (y - y^{(t)})^2$), avec plusieurs étapes on converge toujours (i.e. $\forall \gamma > 0$) vers $(x = 4, y = 1)$: y est constant.

Avec $d = \text{wasser}(W_2^2)$ (le plus petit entre frob et $(x - y^{(t)})^2 + (y - x^{(t)})^2$)

--> en une étape: la limite pour un pas infini est $(x = 4, y = 0)$. On déplace donc le point y qui a une dérivée nulle pour la fonction f à minimiser.

--> plusieurs étapes: quasiment idem, voir les schémas

7.1 DÉTAIL WASSER

Préreq(partie suivante): w_2 avec mesures somme de diracs uniforme -> pb de permutation.

On trouve facilement que l'assignement est l'identité si $x \leq y$ et on inverse dans l'autre cas (avec deux solutions pour l'égalité). On peut réécrire l'argmin avec la distance de wasserstein:

$$(x^{(t+1)}, y^{(t+1)}) = \arg \min_{(x,y)} \frac{1}{2}(x-4)^2 + \frac{1}{2\gamma} \left[\mathbf{1}_{x \leq y} \left((x - x^{(t)})^2 + (y - y^{(t)})^2 \right) + \mathbf{1}_{x > y} \left((x - y^{(t)})^2 + (y - x^{(t)})^2 \right) \right] \quad (44)$$

Dans Eq. (44), soit on paie le coût de transporter x^t vers x et y^t vers y (si $x^t < y^t$, car c'est le plus efficace dans ce cas), soit on paie le transport de x^0 vers y et y^0 vers x .

7.2 WASSERSTEIN: DÉTAILS DE LA PREMIÈRE ÉTAPE

Pour $\gamma < 1/3$.

la première étape de prox sera identique qu'importe la distance utilisé, car $x^{(1)} \leq y^{(0)}$ (avec un petit pas, il n'est pas possible d'obtenir une solution plus loin que $y^{(0)}$). Eq. (44) devient:

$$(x^{(t+1)}, y^{(t+1)}) = \arg \min_{(x,y)} \frac{1}{2}(x-4)^2 + \frac{1}{2\gamma} \left((x - x^{(t)})^2 + (y - y^{(t)})^2 \right) \quad (45)$$

Et on peut résoudre directement:

$$(x^{(1)}, y^{(1)}) = \left(\frac{1}{\gamma + 1} (4\gamma + x^{(0)}), 1 \right) \quad (46)$$

Dès que $\gamma > 1/3$

on a alors $x^{(1)} > y^{(0)}$ (notre objectif est convexe et a pour solution $4 > y^{(0)}$, donc dès que le pas est assez grand, la solution de l'argmin dépasse $y^{(0)}$) et le coût de la distance wasserstein (qui associe x à $y^{(0)}$ et y à $x^{(0)}$ est plus faible)

$$(x^{(t+1)}, y^{(t+1)}) = \arg \min_{(x,y)} \underbrace{\frac{1}{2}(x-4)^2 + \frac{1}{2\gamma}((x-y^{(t)})^2 + (y-x^{(t)})^2)}_{\Phi_1} \quad (47)$$

On a donc un gradient non nul pour y . On a encore un problème convexe qu'on peut résoudre directement:

$$\frac{\partial \Phi_1}{\partial x} = x \left(1 + \frac{1}{\gamma}\right) - \frac{1}{\gamma} - 4 \quad (48)$$

$$\frac{\partial \Phi_1}{\partial y} = \frac{1}{\gamma} y \quad (49)$$

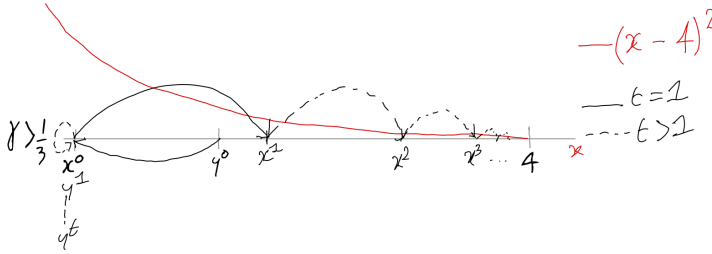
En posant $\nabla \Phi_1 = 0$ on obtient une unique solution (Φ_1 convexe)

$$(x^{(1)}, y^{(1)}) = \left(1 + \frac{3}{1 + \frac{1}{\gamma}}, 0\right) \quad (50)$$

Plusieurs étapes

Résumé: Après une étape de prox, on reset les "positions initiales". Une fois que $x^{(t)} > y^{(t)}$, le mapping sera l'identité (les deux distances produiront les mêmes itérés).

Cas 2. $\gamma > \frac{1}{3}$

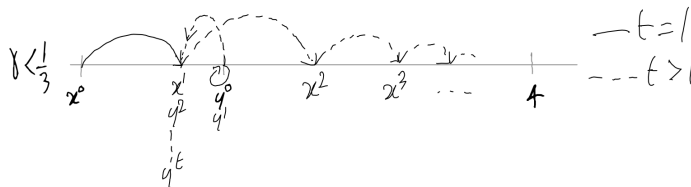


On reprend Eq. (44) avec $t = 1$ pour obtenir $(x, y)^{(t+1=2)}$. Les "positions initiales" sont maintenant $(x, y)^{(1)}$ calculé à la première étape:

$$(x^{(1)}, y^{(1)}) = \left(1 + \frac{3}{1 + \frac{1}{\gamma}}, 0\right) \quad (51)$$

L'objectif étant convexe, l'optimal est toujours à droite de $x^{(1)}$, donc on a toujours $x > y^{(1)} = 0$.

Cas 1. $\gamma < \frac{1}{3}$



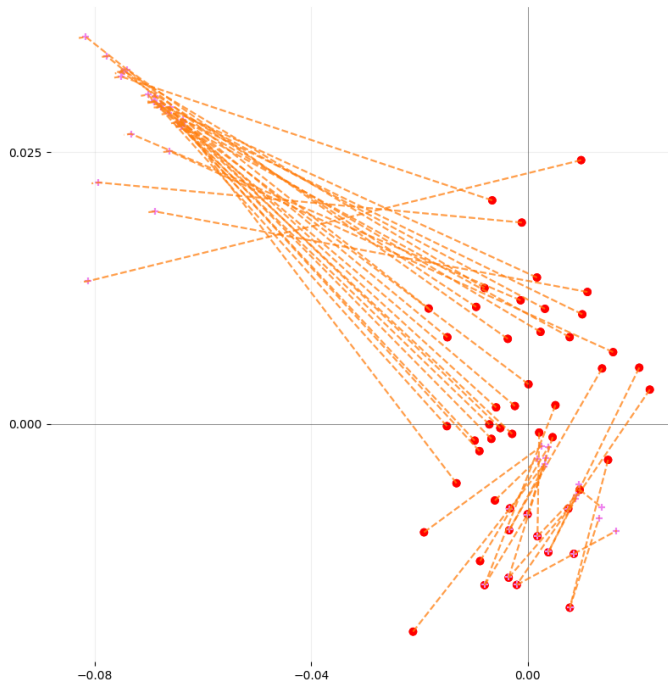
y est identique, x se déplace vers 4. A un moment t , on aura $x^{t+1} > y^0$ ($\forall t' < t, y^{t'} = y^0$) et on peut refaire les calculs de la première étape dans le cas $\gamma > 1/3$ (on a juste x différent).

7.3 EXTENSIONS

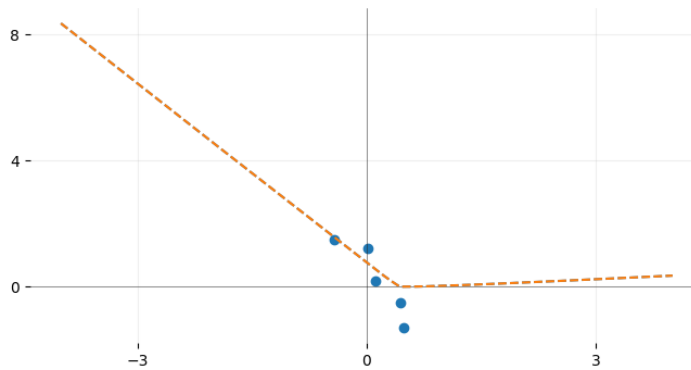
We can project the whole example into the first dimension of \mathbb{R}^2 and rotate it and we would find the same behavior.

Taking a ReLU network loss and weights, it seems unlikely that a problem with two neurons will have a *different output* at convergence with the wasserstein distance.

However with many neurons we can still observe on the graph below, that in a single step of prox there are many exchanges being made, but in the end the two proxs will converge to the same solution.



This simulation is simply some random data that has two obvious directions:



To remove any suspicion we only considered positives neurons here. Indeed all previous simulations have allowed exchanges between positives neurons (that have +1 for the second layer and thus positive outputs) and negative neurons.

It is unclear if I should spend more time formalizing an example where we have an exchange.

7.4 NEXT STEPS

A 'large' stepsize for the proximal point using wasserstein might always happen in some scenarios with many neurons and a few obvious directions.

In practice it is not easy to see the difference in the converged network, in large neurons scenarios where the loss usually reach zero for both distances.

A large scale scenario will be ran (maybe many neurons in a grid again), with pos and neg distances computed differently (requires some code change).

8 INTUITION ON SMALL VS BIG INITIALIZATION IN ReLU

Why is it that as the scale gets smaller, the training dynamic consist of an alignment phase and then a convergence phase?

Notation: Take a neuron $w \in \mathbb{R}^d$. The norm of the neuron is $\|w\|_2 = \sqrt{\sum_{i=1}^d w_i^2}$, the direction of a neuron is a vector of norm 1: $\frac{w}{\|w\|_2}$.

A training step is approximately $w^{(t+1)} = w^{(t)} + \gamma \mathbf{v}$ with γ a coefficient correlated with the current (t) output, labels, error and step size. However, it is not directly correlated to the norm of the neuron.

$\mathbf{v} \in \mathbb{R}^d$ is a partial sum of training examples. It has its own norm and direction.

If $\|w^{(0)}\|$ is large, $w^{(1)}$ will be close to $w^{(0)} + \gamma \mathbf{v}$.

If $\|w^{(0)}\|$ is close to 0, $w^{(1)}$ will be close to equal to $\gamma \mathbf{v}$ as every coefficient of w is negligible compared to the update. Therefore, its direction will (after some updates) be dominated by \mathbf{v} . The same for every neuron that activate the same data points. (Since activation pattern is entirely decided by direction, activation pattern will all converge to extremal vectors..)

Refs: (expés, some results) (Maennel et al., 2018), (the orthogonal paper) (Boursier et al., 2022). (incremental learning) (Berthier, 2022), (scaling path) (Neumayer et al., 2023).

9 JKO

What we compute by using the entropic JKO flow iterations.

$$\forall t > 0, p_{t+1} := \text{Prox}_{\tau f}^{W_\gamma}(p_t) \quad (52)$$

$$= \arg \min_{p \in \text{simplex}} W_\gamma(p, q) + \tau f(p) \quad (53)$$

$$= \arg \min_{p \in \text{simplex}} \left(\min_{\pi \in \Pi(p, q)} \langle c, \pi \rangle + \gamma E(\pi) \right) + \tau f(p) \quad (54)$$

9.1 JKO STEPPING WITH DYKSTRA'S ALGORITHM

--> [Entropic WGF 2015](#)

$$p_{t+1} := \text{Prox}_{\tau f}^{W_\gamma}(p_t) \quad (55)$$

$$= \arg \min_{p \in \text{simplex}} W_\gamma(p, q) + \tau f(p) \quad (56)$$

$$= \arg \min_{p \in \text{simplex}} \left(\min_{\pi \in \Pi(p, q)} \langle c, \pi \rangle + \gamma E(\pi) \right) + \tau f(p) \quad (57)$$

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 formally discretize the above mentioned 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

```

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

```

- [Meta Optimal Transport \(paper\)](#) and [\(code git\)](#): InputConvexNN to predict solution of OT problem
- [JKOnet \(paper\)](#) and [\(code git\)](#):
 - [/models](#) -> sinkhorn loss defined in loss.py, differentiable loop in fixed point.py
 - next step: trying to create the right [Geometry](#) object from OTT library, which is what's used for sinkhorn

9.2 PAPERS

Paper with a [specific case that doesn't match ours](#):

[large-scale wasserstein gradient flows](#) about computing WGF. They discretize over time, but not exactly space. They reformulate so that the argmin is over convex functions, and then discretize cvx functions using ICNN.

REFERENCES

- Francis Bach and Lenaïc Chizat. Gradient descent on infinitely wide neural networks: Global convergence and generalization. *arXiv preprint*. doi: 10.48550/arXiv.2110.08084. URL <http://arxiv.org/abs/2110.08084>.
- Raphael Berthier. Incremental learning in diagonal linear networks. *arXiv preprint arXiv:1803.08367*, 2022. URL <https://jmlr.org/papers/volume24/22-1395/22-1395.pdf>.
- Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *arXiv preprint arXiv:1803.08367*, 2022. doi: 10.48550/arXiv.2206.00939. URL <http://arxiv.org/abs/2206.00939>.
- Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *arXiv preprint*. URL <http://arxiv.org/abs/1512.02783>.
- Lénaïc Chizat and Francis R. Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Adv. Neural Inf. Process. Syst. 31 Annu. Conf. Neural Inf. Process. Syst. 2018 NeurIPS 2018 Dec. 3-8 2018 Montr. Can.*, pp. 3040–3050. URL <https://proceedings.neurips.cc/paper/2018/hash/a1afc58c6ca9540d057299ec3016d726-Abstract.html>.
- Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.
- Sebastian Neumayer, Lénaïc Chizat, and Michael Unser. On the effect of initialization: The scaling path of 2-layer neural networks. *arXiv preprint arXiv:1803.08367*, 2023. URL <http://arxiv.org/abs/2303.17805>.