# *p*-voltages: Laplacian Regularization for Semi-Supervised Learning on High-Dimensional Data

Nick Bridle, Xiaojin Zhu

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

## SETTING

- High-dimensional i.i.d. data
- We want to compute a "smooth" function over the data for classification or regression
- Semi-supervised setting - small number of labeled data points, large number of unlabeled data points
- We construct a similarity graph over the data
  - e.g. knn-graph or epsilon graph
  - edge weights are some Euclidean similarity kernel

## PROBLEM WITH EXISTING METHOD

Existing method: Laplacian Regularization algorithm

- Idea: label the nodes in a graph with a harmonic function
- Corresponds to the voltages in an electric network
- Hold labeled nodes fixed with v=1 (positive) or v=0 (negative)

$$\min_v \left\{ \sum_{(a,b)\in E} w_{ab}(v_a - v_b)^2 \;\middle|\; v_s - v_t = 1 \right\}$$

Problem: Leads to a "flat, spiky" pathology when the number of points grows to infinity

- Almost-everywhere constant function: "flatness"
- In local neighborhoods around labeled points, dramatic change in voltage: "spikiness"
- From an electricity perspective, current is distributed too widely to result in significant voltage drop in most of the graph
- Not suitable for classification or regression in high dimensions or for large datasets

## p-ELECTRIC NETWORKS AND THE p-VOLTAGES ALGORITHM

p-electricity is hypothetical form of electricity which concentrates itself on fewer paths. The concentration parameter is p.

- Our world: $p = 2$
- We consider $1 < p < p^* := \frac{d}{d-1}$, where d is the dimension of the underlying space

Voltages and currents follow p-Ohm's Law in a p-electric network:

$$v_a - v_b = sign(i_{ab})|i_{ab}|^{p-1} r_{ab}$$

The p-voltages algorithm uses this modified optimization problem:

$$\min_v \left\{ \sum_{(a,b)\in E} \frac{|v_a - v_b|^{\frac{p}{p-1}}}{r_{ab}^{\frac{1}{p-1}}} \;\middle|\; v_s - v_t = 1 \right\}$$

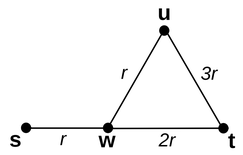
We can use a standard numerical solver (e.g. MATLAB Optimization Toolbox) to calculate this. It is a convex objective function over a convex set, so a unique solution exists.

## OUR CONTRIBUTIONS

We disprove a conjecture from [1] which proposes that p-voltages can be used as a computationally convenient way to classify by p-resistances. That is, for all p > 1,

$$v_u^* - v_t^* > v_s^* - v_u^* \iff R_p(u,t) > R_p(s,u)$$

To do this, we analyze a simple counter-example graph (shown below) and note that in numerical experiments, the property often does not hold for all nodes in the graph. However, we still advocate using the p-voltages algorithm in its own right for classification.




We prove that p-voltage solutions are well-formed with two theorems:

Theorem 1.
If we define the following quantities,

$$V_s := \max\{|v_s - v_u|, u \in \mathcal{N}(s)\}$$
$$V_t := \max\{|v_u - v_t|, u \in \mathcal{N}(t)\}$$
$$V := v_s - v_t = R_p(s,t).$$

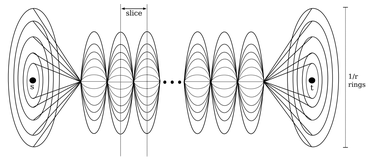
then,

$$\frac{V_s + V_t}{V} \to 0 \text{ as } n \to \infty.$$

That is, the maximum voltage drop in the local neighborhood of the source and sink points is negligible as the size of the graph grows. Thus, the solution does not have the spike problem.

[1] M. Alamgir and U. von Luxburg. *Phase transition in the family of p-resistances*. NIPS 2011.

Theorem 2.
Consider the ratio $\frac{V_M}{V}$, which is the portion of the total graph-wide change in p-voltage which occurs in region M. If M is a *substantial region* (defined in the paper) for constant $c \in (0,1)$, and $p < \frac{d}{d-1}$, then we know that $\frac{V_M}{V} \geq c$.

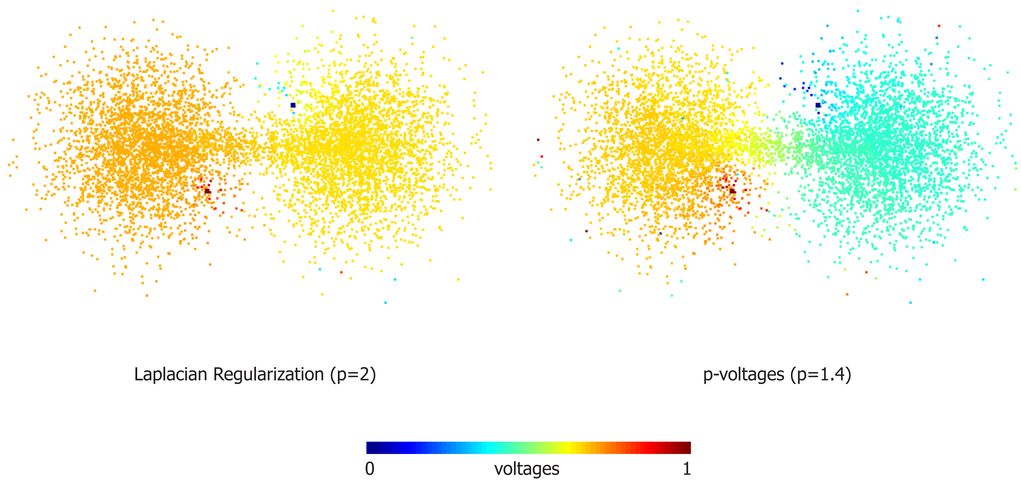



Contracted graph used to lower bound the p-resistance of the graph in the proof of Theorem 2.

Using this theorem we can construct several different substantial regions in a graph and conclude that the voltage drop must be distributed throughout. Thus, the solution is not flat, either.

## ILLUSTRATIVE EXAMPLE

A 3-d "barbell" graph is formed with two Gaussians connected by a cylinder. The voltages of a single labeled source node (red) and sink node (blue) are held fixed. The resulting voltages are shown as a spectrum from red (1) to blue (0).




*This research is supported in part by National Science Foundation grant IIS-0916038.*

## EMPIRICAL RESULTS

We compared the p-voltages algorithm to the standard Laplacian Regularization algorithm as well as a version which uses the ad-hoc Class Mass Normalization (CMN) heuristic. We also compared with the Iterated Laplacian algorithm, which attempts to solve the problem using a higher-order regularizer.

The datasets come from the popular Chapelle Semi-Supervised Learning benchmark, and contain both synthetic and "real-world" datasets. Results are averaged over twelve training-test set splits.

| Dataset | LapReg | LapReg + CMN | IterLap | p-voltages |
|---|---|---|---|---|
| g241c | 48.95 ± 4.38 | 22.46 ± 1.42 | 19.40 ± 4.88 | 33.74 ± 7.20 |
| g241n | 49.93 ± 1.20 | 30.88 ± 3.43 | 13.15 ± 0.97 | 29.91 ± 3.67 |
| Digit1 | 8.81 ± 0.56 | 3.74 ± 1.05 | 2.24 ± 0.81 | 3.14 ± 0.95 |
| USPS | 19.19 ± 0.67 | 10.91 ± 1.06 | 4.58 ± 0.86 | 7.54 ± 1.98 |
| BCI | 46.89 ± 2.33 | 46.19 ± 2.14 | 45.67 ± 2.75 | 45.03 ± 2.78 |

Average Test Set Classification Error Percentage & Standard Error on Chapelle Benchmark

p-voltages outperforms Laplacian Regularization on all but one of the datasets. However, empirically the algorithm does not outperform the state-of-the-art Iterated Laplacian algorithm.



contacts: {nbridle, jerryzhu}@cs.wisc.edu