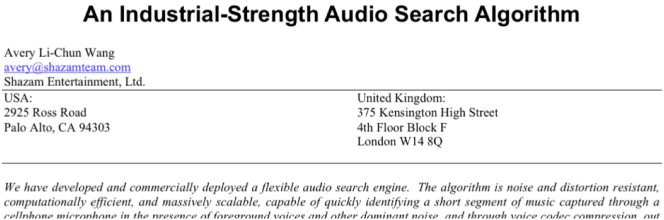
Network Flows in Semi-Supervised Learning via Total Variation Minimization

Alexander Jung
Dept. of Computer Science, Aalto University

April 9, 2019

I like Kill Bill

- watched Kill Bill recently
- fighting scence with a cool background song
- smartphone dug out the title in seconds!
- song completely unrelated to my preferences



An Industrial-Strength Audio Search Algorithm

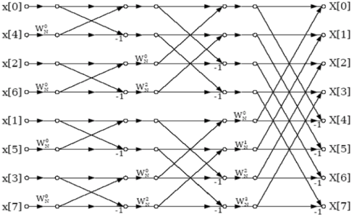
Avery Li-Chun Wang
avery@shazamteam.com
Shazam Entertainment, Ltd.

USA:
2925 Ross Road
Palo Alto, CA 94303

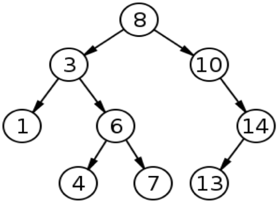
United Kingdom:
375 Kensington High Street
4th Floor Block F
London W14 8Q

We have developed and commercially deployed a flexible audio search engine. The algorithm is noise and distortion resistant, computationally efficient, and massively scalable, capable of quickly identifying a short segment of music captured through a cellphone microphone in the presence of foreground voices and other dominant noise, and through voice codec compression, out

fast Fourier transform



fast search



Outline

Big Data Fuels Machine Learning

availability of vast amounts of training data allows

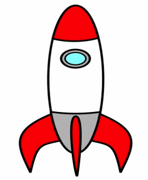
to train extremely complex models such as

deep neural networks

Andrew Ng's Rocket Picture

Big Data

+



Complex Model

=

Modern AI/ Deep Learning

AI Everywhere

- Shazam identifies the ear-worm tune you are listening to
- spam filters keep your inbox tidy
- Google.com became personal Jeannie

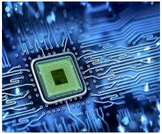
A Key Principle

organize data and computation
as networks

Big Data over Networks

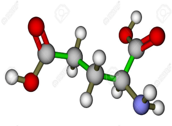
datasets and models often have intrinsic network structure

chip design



internet

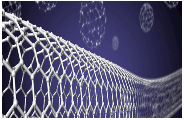
bioinformatics



social networks

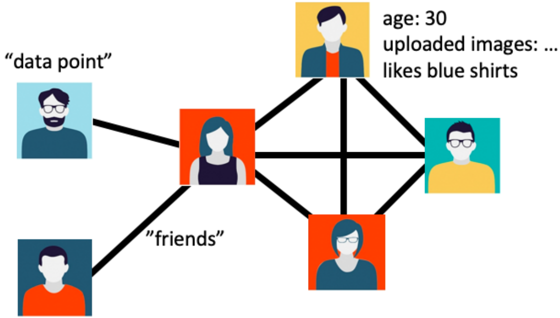
universe

material science



cf. L. Lovász, "Large Networks and Graph Limits"

Partially Labeled Networked Data



- data points have features (posted videos, links, texts,. . .) and labels (preference for certain product)
- features are “cheap”, labels are “costly”
- try to get along with few labels

Graph Signals in Networked Data

- networked data with empirical graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$
- similar data points i, j connected by edge $\{i, j\} \in \mathcal{E}$
- level of similarity quantified by weight $W_{i,j} > 0$
- data points $i \in \mathcal{V}$ characterized by labels x_i
- predictor/classifier maps nodes to predicted label $\hat{x}_i$
- represent predictor by graph signal $\hat{\mathbf{x}} = \big(\hat{x}_1, \ldots, \hat{x}_N\big)^T$

Empirical Loss Minimization over Graphs

- acquire labels for few data points in training set $\mathcal{M}$
- learn entire labelling $\mathbf{x}$ using network $\mathbf{W}$ and labels on $\mathcal{M}$
- aim at small empirical (training) error $\sum_{i \in \mathcal{M}} f_i(\hat{x}_i; x_i)$
- loss function $f_i(\cdot)$ for data point $i \in \mathcal{V}$
- e.g., $f_i(\ldots) := (\hat{x}_i - x_i)^2$ or $f_i(\ldots) := 1/(1 + \exp(-x_i \hat{x}_i))$

Clustering Hypothesis

- well-connected subsets (clusters) of $\mathcal{G}$ have similar labels x_i
- amounts to requiring small total variation (TV)

$$\|\mathbf{x}\|_{\mathrm{TV}} := \sum_{\{i,j\} \in \mathcal{E}} W_{i,j} |x_i - x_j|$$

- using (weighted) incidence matrix $\mathbf{D}$ of empirical graph,

$$\|\mathbf{x}\|_{\mathrm{TV}} = \|\mathbf{D}\mathbf{x}\|_1$$

Outline

The Learning Problem

- observe initial labels x_i for few data points $i \in \mathcal{M}$ $(\ll \mathcal{V})$
- aim at learning all labels x_i, for $i \in \mathcal{V}$
- empirical risk incurred by particular hypothesis $\hat{\mathbf{x}}$ is

$$\mathcal{E}(\hat{\mathbf{x}}) := \sum_{i \in \mathcal{M}} f_i(\hat{x}_i; x_i)$$

  with some loss function $f_i(\cdot;\cdot) \in \mathbb{R}$ associated with node i

- balance empirical risk $\mathcal{E}(\hat{\mathbf{x}})$ with TV $\|\hat{\mathbf{x}}\|_{\mathrm{TV}}$

TV Minimization - The Network Lasso

- network Lasso (nLasso) [Hallac, 2015] formulates TV min as

$$\min_{\hat{\mathbf{x}}} \sum_{i \in \mathcal{M}} f_i(\hat{x}_i; x_i) + \lambda \|\hat{\mathbf{x}}\|_{\mathrm{TV}}$$

- large λ enforces small total variation $\|\hat{\mathbf{x}}\|_{\mathrm{TV}}$
- small λ enforces small empirical error
- we can enforce consistency with initial labels by using

$$f_i(\hat{x}, x) = \infty \text{ for } x \neq \hat{x}$$

Structure of Network Lasso

- nLasso has particular structure:

$$\min_{\hat{\mathbf{x}}} \sum_{i \in \mathcal{M}} f_i(\hat{x}_i; x_i) + \lambda \|\hat{\mathbf{x}}\|_{\mathrm{TV}}$$

- sum of two non-smooth convex components
- minimizing each component individually is easy

Convex Optimization and Fixed Points

- nLasso: $\min_{\hat{\mathbf{x}}} \sum_{i \in \mathcal{M}} f_i(\hat{x}_i; x_i) + \lambda \|\hat{\mathbf{x}}\|_{\mathrm{TV}}$
- objective sum of two non-smooth convex components
- nLasso delivers $\hat{\mathbf{x}}$ if and only if $\mathbf{0} \in \partial f(\hat{\mathbf{x}})$
- rewrite $\mathbf{0} \in \partial f(\hat{\mathbf{x}})$ as $\hat{\mathbf{x}} = \mathcal{P}\hat{\mathbf{x}}$ with some operator $\mathcal{P}$
- different options for $\mathcal{P}$ (EXPLOIT THIS FREEDOM!)
- well-known methods such as ADMM obtained for particular $\mathcal{P}$

Iterative Methods for TV Minimization

- TV min characterized by $\hat{\mathbf{x}} = \mathcal{P}\hat{\mathbf{x}}$
- compute $\hat{\mathbf{x}}$ by fixed-point iteration $\hat{\mathbf{x}}^{(k+1)} = \mathcal{P}\hat{\mathbf{x}}^{(k)}$
- scalable implementation via message passing
- primal-dual method is fixed-point iteration for particular $\mathcal{P}$ obtained from convex duality

The Dual of TV Min

- TV min. is primal problem $\min_{\tilde{\mathbf{x}}} g(\mathbf{D}\tilde{\mathbf{x}}) + h(\tilde{\mathbf{x}})$
- with $g(\mathbf{y}) := \|\mathbf{y}\|_1$, and $h(\tilde{\mathbf{x}}) := \begin{cases} \infty \text{ if } \tilde{x}_i \neq x_i \text{ for } i \in \mathcal{M} \\ 0 \text{ else.} \end{cases}$
- define dual problem $\max_{\mathbf{y} \in \mathbb{R}^{\mathcal{E}}} -h^*(-\mathbf{D}^T\mathbf{y}) - g^*(\mathbf{y})$
- $h^*(\mathbf{x}), g^*(\mathbf{y})$ are convex conjugates of $h(\mathbf{x}), g(\mathbf{y})$
- $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ primal-dual optimal if and only if

$$-(\mathbf{D}^T\hat{\mathbf{y}}) \in \partial h(\hat{\mathbf{x}}),\ \mathbf{D}\hat{\mathbf{x}} \in \partial g^*(\hat{\mathbf{y}})$$

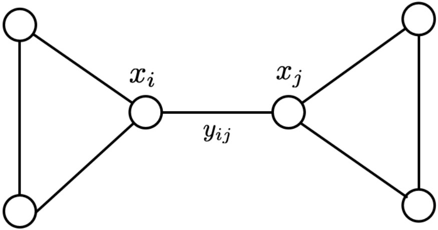
- entries y_{ij} of dual $\mathbf{y}$ are signal values on edges $\{i,j\} \in \mathcal{E}$

Convex Conjugates

Primal-Dual Method for TV Min

- construct coupled sequences $\hat{\mathbf{x}}^{(k)}$ and $\hat{\mathbf{y}}^{(k)}$ for $k = 0, 1, \ldots$
- asymptotic optimal $\lim_{k \to \infty} \hat{\mathbf{x}}^{(k)} \rightarrow \hat{\mathbf{x}}$, $\lim_{k \to \infty} \hat{\mathbf{y}}^{(k)} \rightarrow \hat{\mathbf{y}}$
- update $\hat{\mathbf{x}}^{(k)}, \hat{\mathbf{y}}^{(k)} \mapsto \hat{\mathbf{x}}^{(k+1)}, \hat{\mathbf{y}}^{(k+1)}$ uses only local computation
- amounts to message passing on empirical graph $\mathcal{G}$
- see `https://arxiv.org/abs/1901.09838` for details

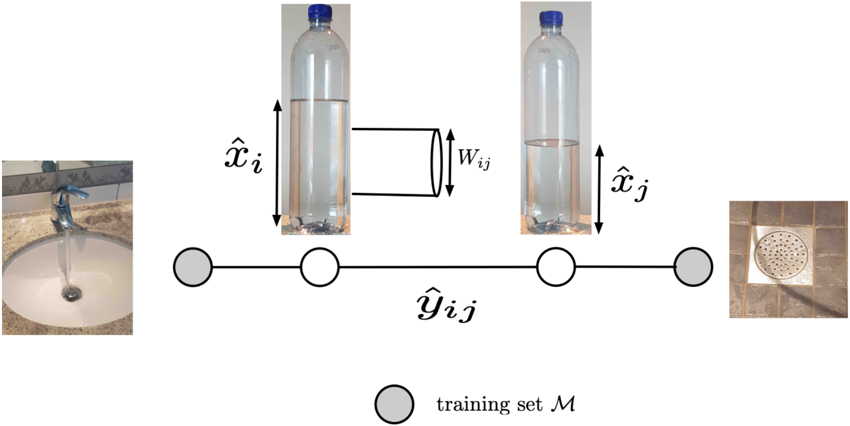
Primal-Dual Message Passing



The Dual of TV Min is Maximum Flow

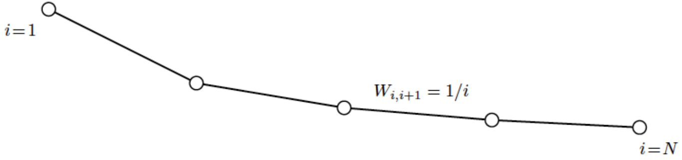
- TV min. dual $\max_{\mathbf{y} \in \mathbb{R}^{\mathcal{E}}} -h^{*}(-\mathbf{D}^{T}\mathbf{y}) - g^{*}(\mathbf{y})$
- equivalent to maximizing a flow on empirical graph
- flow on edge $\{i,j\} \in \mathcal{E}$ is $f_{ij} = y_{ij} W_{ij}$
- dual solution $\hat{\mathbf{y}}$ of TV min characterized by
  - $|\hat{f}_{ij}| = |\hat{y}_{ij} W_{ij}| \leq W_{ij}$ (capacity constraints)
  - $\sum_{j \in \mathcal{N}(i)} \hat{f}_{ij} = 0$ for $i \notin \mathcal{M}$ (flow conservation)
  - $\sum_{i \in \mathcal{M}} x_i \sum_{j \in \mathcal{N}(i)} \hat{f}_{ij}$ is maximal

message passing formulation of primal-dual method for TV min provides a distributed max. flow algorithm !



Computational Complexity of TV Min

- it can be shown that $\|\hat{\mathbf{x}}^{(k)}\|_{\mathrm{TV}} - \|\mathbf{x}\|_{\mathrm{TV}} \propto 1/k$
- "$1/k$" convergence is optimal without further assumptions
- cannot be overcome by any method for chain graph

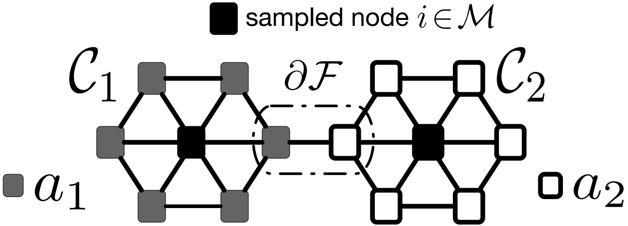


Outline

- implement clustering hypothesis by piece-wise constant signals

$$x_i = \sum_{\mathcal{C} \in \mathcal{F}} a_{\mathcal{C}} \mathcal{I}_{\mathcal{C}}[i] \quad \text{, with } \mathcal{I}_{\mathcal{C}}[i] = \begin{cases} 1 & \text{if } i \in \mathcal{C} \\ 0 & \text{else.} \end{cases}$$

using partition $\mathcal{F} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{F}|}\}$ with disjoint clusters $\mathcal{C}_l$



Good Clusters - Small Total Variation

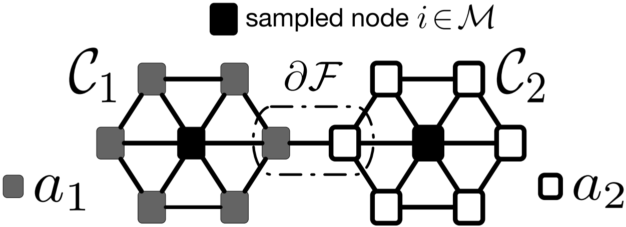
- we allow for arbitrary partition $\mathcal{F} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{|\mathcal{F}|}\}$
- our results are most useful for "reasonable clusters" $\mathcal{C}_l$
- cluster boundary $\partial\mathcal{F}$ with small average weight

$$\sum_{\text{boundary}} W_{i,j} \ll \sum_{\text{interior}} W_{i,j}$$

- for such clusters, piece-wise constant signals have small TV

The Learning (Recovery) Problem

- networked data $\mathcal{G}$ with known labels $x_i \in \mathcal{M}$
- labelling $\mathbf{x}$ is piece-wise constant (clustering assumption)



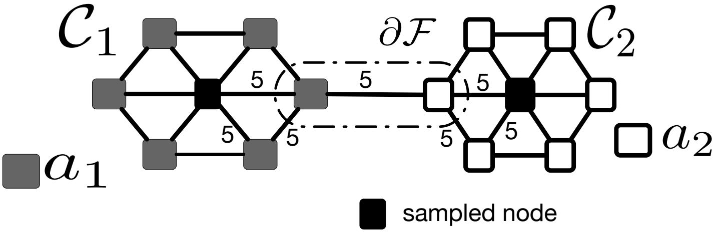
Idea of Nullspace Condition

- stack initial labels into vector $\mathbf{y} \in \mathbb{R}^{\mathcal{M}}$
- recover signal $\mathbf{x}$ from "measurements" $\mathbf{y} = \mathbf{M}\mathbf{x}$
- selector matrix $\mathbf{M}$ with rows $\{\mathbf{e}_i\}_{i \in \mathcal{M}}$
- recovery impossible for any $\mathbf{x}$ in nullspace $\mathcal{K}(\mathbf{M})$ of $\mathbf{M}$
- have to ensure $\mathcal{K}(\mathbf{M}) \cap \{$ piecewise constant signals$\} = \emptyset$
- piece-wise constant signals must not vanish on $\mathcal{M}$

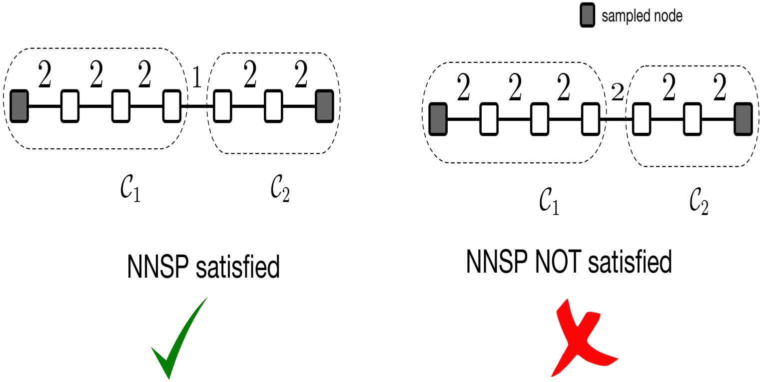
The Network Nullspace Property (NNSP)

- consider partition $\mathcal{F} = \{\mathcal{C}_1, \ldots\}$ of the empirical graph $\mathcal{G}$
- training set $\mathcal{M}$ satisfies network nullspace property w.r.t. $\mathcal{F}$, denoted NNSP-($\mathcal{M}$,$\mathcal{F}$), if there exist flow f_{ij} with

$$f_{ij} = 2W_{i,j} \text{ for } \{i,j\} \in \partial\mathcal{F}$$



When is NNSP Satisfied?



NNSP implies SLP Recovers Clustered Signals

Theorem. Consider networked data with labels $\mathbf{x}$ which can be well approximated as piece-wise constant over a partition $\mathcal{F}$. We have access to the labels only for the data points in the training set $\mathcal{M} \subseteq \mathcal{V}$. Then, if NNSP-$(\mathcal{M}, \mathcal{F})$ holds,

$$\min_{\hat{\mathbf{x}}} \|\hat{\mathbf{x}}\|_{\mathrm{TV}} \text{ s.t. } \hat{x}_i = x_i \text{ for all } i \in \mathcal{M}$$

has a unique solution which coincides with true labels $\mathbf{x}$.

NNSP implies SLP is Robust

Theorem. Consider networked data with labels $\mathbf{x}$ which are piece-wise constant over a partition $\mathcal{F}$. We have access to the labels only for the data points in the training set $\mathcal{M} \subseteq \mathcal{V}$. Then, if NNSP-$(\mathcal{M}, \mathcal{F})$ holds, TV min delivers $\hat{\mathbf{x}}$ with

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_{\mathrm{TV}} \leq 6 \underbrace{\min_{\mathbf{a} \in \mathbb{R}^{|\mathcal{F}|}} \|\mathbf{x} - \sum_{\mathcal{C} \in \mathcal{F}} a_{\mathcal{C}} \mathcal{I}_{\mathcal{C}}[\cdot]\|_{\mathrm{TV}}}_{\text{model misfit}} .$$

Outline

So what...?

- implemented TV min. using primal-dual method
- dual of TV min. is a max. flow problem
- message passing for TV min provides max flow method
- TV min is accurate when suff. large flows exist
- can be extended to node-wise models (instead of labels)

Reading Material

- AJ et.al., "Semi-supervised Learning in Network-Structured Data via Total Variation Minimization", https://arxiv.org/abs/1901.09838
- AJ, N. Tran, "Localized Linear Regression in Networked Data", https://arxiv.org/abs/1903.11178
- D. Bertsekas, "Network Optimization: Continuous and Discrete Models", Athena Scientific, 1998

Thats it Folks!