Graphs in Machine Learning

Michal Valko

*Inria Lille - Nord Europe, France*

TA: Daniele Calandriello

Partially based on material by: Branislav Kveton, Partha Niyogi, Rob Fergus
Last Lecture

- Inductive and transductive semi-supervised learning
- Manifold regularization
- Theory of Laplacian-based manifold methods
- Transductive learning stability based bounds
- Online Semi-Supervised Learning
- Online incremental $k$-centers
This Lecture

- Examples of applications of online SSL
- Analysis of online SSL
- SSL Learnability
- When does graph-based SSL provably help?
- Scaling harmonic functions to millions of samples
Previous Lab Session

- 14. 11. 2016 by Daniele Calandriello
- Content
  - Semi-supervised learning
  - Graph quantization
  - Offline face recognizer
- Install VM (in case you have not done it yet for TD1)
- Short written report
- Questions to piazza
- **Deadline: 28. 11. 2016**
Next Lab Session/Lecture

- 28. 11. 2016 by Daniele.Calandriello@inria.fr
- Content (this time lecture in class + coding at home)
  - Large-scale graph construction and processing (in class)
  - Scalable algorithms:
    - Online face recognizer (to code in Matlab)
    - Iterative label propagation (to code in Matlab)
    - Graph sparsification (presented in class)
- AR: record a video with faces
- Short written report
- Questions to piazza
- **Deadline: 12. 12. 2016**

- [http://researchers.lille.inria.fr/~calandri/teaching.html](http://researchers.lille.inria.fr/~calandri/teaching.html)
Final Class projects

- detailed description on the class website
- preferred option: you come up with the topic
- theory/implementation/review or a combination
- one or two people per project (exceptionally three)
- grade 60%: report + short presentation of the team
- deadlines
  - 21. 11. 2016 - recommended DL for taking projects Today!
  - 28. 11. 2016 - hard DL for taking projects
  - 05. 01. 2017 - submission of the project report
  - 09. 01. 2017 or later - project presentation
- list of suggested topics on piazza
Online SSL with Graphs

Video examples

http://www.bkveton.com/videos/Coffee.mp4

http://www.bkveton.com/videos/Ad.mp4

http://researchers.lille.inria.fr/~valko/hp/serve.php?
what=publications/kveton2009nipsdemo.adaptation.mov

http://researchers.lille.inria.fr/~valko/hp/serve.php?
what=publications/kveton2009nipsdemo.officespace.mov

http://bcove.me/a2derjeh

SSL with Graphs: Some experimental results

- 8 people classification
- Making funny faces
- 4 faces/person are labeled

![Image of a video frame with graphical analysis](Image.png)
SSL with Graphs: Some experimental results

- One person moves among various indoor locations
- 4 labeled examples of a person in the cubicle

Online HFS outperforms OSSB (even when the weak learners are chosen using future data)

Online HFS yields better results than a commercial solution at 20% of the computational cost
SSL with Graphs: Some experimental results

- **Logging in** with faces instead of password
- Able to **learn** and improve
SSL with Graphs: Some experimental results

- 16 people log twice into a tablet PC at 10 locations

Online HFS yields better results than a commercial solution at 20% of the computational cost
Online SSL with Graphs: Analysis

What can we guarantee?

Three sources of error

- generalization error — if all data: \((\ell_t^* - y_t)^2\)
- online error — data only incrementally: \((\ell^o_t[t] - \ell_t^*))^2\)
- quantization error — memory limitation: \((\ell^q_t[t] - \ell^o_t[t])^2\)

All together:

\[
\frac{1}{N} \sum_{t=1}^{N} (\ell^q_t[t] - y_t)^2 \leq \frac{9}{2N} \sum_{t=1}^{N} (\ell_t^* - y_t)^2 + \frac{9}{2N} \sum_{t=1}^{N} (\ell^o_t[t] - \ell_t^*)^2 + \frac{9}{2N} \sum_{t=1}^{N} (\ell^q_t[t] - \ell^o_t[t])^2
\]

Since for any \(a, b, c, d \in [-1, 1]\):

\[
(a - b)^2 \leq \frac{9}{2} \left[ (a - c)^2 + (c - d)^2 + (d - b)^2 \right]
\]
Online SSL with Graphs: Analysis

Bounding transduction error \((\ell_t^* - y_t)^2\)

If all labeled examples \(l\) are i.i.d., \(c_l = 1\) and \(c_l \gg c_u\), then

\[
R(\ell^*) \leq \hat{R}(\ell^*) + \beta + \sqrt{\frac{2 \ln(2/\delta)}{n_l}} (n_l \beta + 4)
\]

transductive error \(\Delta_T(\beta, n_l, \delta)\)

\[
\beta \leq 2 \left[ \frac{\sqrt{2}}{\gamma_g + 1} + \sqrt{2 n_l} \frac{1 - c_u}{c_u} \frac{\lambda_M(L) + \gamma_g}{\gamma_g^2 + 1} \right]
\]

holds with the probability of \(1 - \delta\), where

\[
R(\ell^*) = \frac{1}{N} \sum_t (\ell_t^* - y_t)^2 \quad \text{and} \quad \hat{R}(\ell^*) = \frac{1}{n_l} \sum_{t \in l} (\ell_t^* - y_t)^2
\]

How should we set \(\gamma_g\)?
Online SSL with Graphs: Analysis

Bounding online error \((\ell^o_t[t] - \ell^*_t)^2\)

Idea: If \(L\) and \(L^o\) are regularized, then HFSs get closer together.

since they get closer to zero

Recall \(\ell = (C^{-1}Q + I)^{-1}y\), where \(Q = L + \gamma gI\)

and also \(v \in \mathbb{R}^{n \times 1}, \lambda_m(A)\|v\|_2 \leq \|Av\|_2 \leq \lambda_M(A)\|v\|_2\)

\[
\|\ell\|_2 \leq \frac{\|y\|_2}{\lambda_m(C^{-1}Q + I)} = \frac{\|y\|_2}{\frac{\lambda_m(Q)}{\lambda_M(C)} + 1} \leq \frac{\sqrt{n_l}}{\gamma_g + 1}
\]

Difference between offline and online solutions:

\[
(\ell^o_t[t] - \ell^*_t)^2 \leq \|\ell^o_t[t] - \ell^*_t\|_\infty^2 \leq \|\ell^o_t[t] - \ell^*_t\|_2^2 \leq \left(\frac{2\sqrt{n_l}}{\gamma_g + 1}\right)^2
\]

Again, how should we set \(\gamma_g\)?
Online SSL with Graphs: Analysis

Bounding quantization error \((\ell_q^t[t] - \ell_o^t[t])^2\)

How are the quantized and full solution different?

\[
\ell^* = \min_{\ell \in \mathbb{R}^N} (\ell - y)^T C (\ell - y) + \ell^T Q \ell
\]

In Q! \(Q^o\) (online) vs. \(Q^q\) (quantized)

We have: \(\ell^o = (C^{-1}Q^o + I)^{-1}y\) vs. \(\ell^q = (C^{-1}Q^q + I)^{-1}y\)

Let \(Z^q = C^{-1}Q^q + I\) and \(Z^o = C^{-1}Q^o + I\).

\[
\ell^q - \ell^o = (Z^q)^{-1}y - (Z^o)^{-1}y = (Z^qZ^o)^{-1}(Z^o - Z^q)y
\]

\[
= (Z^qZ^o)^{-1}C^{-1}(Q^o - Q^q)y
\]
Online SSL with Graphs: Analysis

Bounding quantization error \((\ell_t^q[t] - \ell_t^o[t])^2\)

\[
\ell^q - \ell^o = (Z^q)^{-1}y - (Z^o)^{-1}y = (Z^qZ^o)^{-1}(Z^o - Z^q)y \\
= (Z^qZ^o)^{-1}C^{-1}(Q^o - Q^q)y
\]

\[
\|\ell^q - \ell^o\|_2 \leq \frac{\lambda_M(C^{-1})\|Q^q - Q^o\|y\|_2}{\lambda_m(Z^q)\lambda_m(Z^o)}
\]

\(\|\cdot\|_F\) and \(\|\cdot\|_2\) are compatible and \(y_i\) is zero when unlabeled:

\[
\|(Q^q - Q^o)y\|_2 \leq \|Q^q - Q^o\|_F \cdot \|y\|_2 \leq \sqrt{n_l}\|Q^q - Q^o\|_F
\]

Furthermore, \(\lambda_m(Z^o) \geq \frac{\lambda_m(Q^o)}{\lambda_M(C)} + 1 \geq \gamma_g\) and \(\lambda_M(C^{-1}) \leq c_u^{-1}\)

We get \(\|\ell^q - \ell^o\|_2 \leq \frac{\sqrt{n_l}}{c_u\gamma_g^2}\|Q^q - Q^o\|_F\)
Online SSL with Graphs: Analysis

Bounding quantization error \((\ell_t^q[t] - \ell_t^o[t])^2\)

The quantization error depends on \(\|Q^q - Q^o\|_F = \|L^q - L^o\|_F\).

When can we keep \(\|L^q - L^o\|_F\) under control?

Charikar guarantees distortion error of at most \(Rm/(m - 1)\)

For what kind of data \(\{x_i\}_{i=1,...,n}\) is the distortion small?

Assume manifold \(\mathcal{M}\)

- all \(\{x_i\}_{i \geq 1}\) lie on a smooth \(s\)-dimensional compact \(\mathcal{M}\)
- with boundary of bounded geometry Def. 11 of Hein [HAL07]
  - should not intersect itself
  - should not fold back onto itself
  - has finite volume \(V\)
  - has finite surface area \(A\)
Bounding quantization error \((\ell^q_t[t] - \ell^o_t[t])^2\)

Bounding \(\|L^q - L^o\|_F\) when \(x_i \in \mathcal{M}\)

Consider \(k\)-sphere packing of radius \(r\) with centers contained in \(\mathcal{M}\).

What is the maximum volume of this packing?

\[ kc_s r^s \leq V + A c_M r \text{ with } c_s, c_M \text{ depending on dimension and } \mathcal{M}. \]

If \(k\) is large \(\rightarrow r < \text{injectivity radius}\) of \(\mathcal{M}\) [HAL07] and \(r < 1\):

\[ r < \left( \frac{V + A c_M}{kc_s} \right)^{1/s} = \mathcal{O} \left( k^{-1/s} \right) \]

\(r\)-packing is a \(2r\)-covering:

\[ \max_{i=1,\ldots,N} \|x_i - c\|_2 \leq R m / (m - 1) \leq 2(1 + \varepsilon) \mathcal{O} \left( k^{-1/s} \right) = \mathcal{O} \left( k^{-1/s} \right) \]

But what about \(\|L^q - L^o\|_F\)?
Online SSL with Graphs: Analysis

Bounding quantization error \((\ell^q_t[t] - \ell^o_t[t])^2\)

If similarity is \(M\)-Lipschitz, \(L\) is normalized,

\[
c^o_{ij} = \sqrt{D^o_{ii}D^o_{jj}} > c_{\text{min}}N:
\]

\[
L^q_{ij} - L^o_{ij} = \frac{W^q_{ij}}{c^q_{ij}} - \frac{W^o_{ij}}{c^o_{ij}}
\]

\[
\leq \frac{W^q_{ij} - W^o_{ij}}{c^q_{ij}} + \frac{W^q_{ij}(c^q_{ij} - c^o_{ij})}{c^o_{ij}c^q_{ij}}
\]

\[
\leq \frac{4MRm}{(m-1)c_{\text{min}}N} + \frac{4M(NMRm)}{((m-1)c_{\text{min}}N)^2}
\]

\[
= O \left( \frac{R}{N} \right)
\]

Finally, \(\|L^q - L^o\|_F^2 \leq N^2O(R^2/N^2) = O(k^{-2/s})\).

Are the assumptions reasonable?
Online SSL with Graphs: Analysis

Bounding quantization error \((\ell^q_t[t] - \ell^o_t[t])^2\)

We showed \(\|L^q - L^o\|_F^2 \leq N^2 O(R^2/N^2) = O(k^{-2/s}) = O(1)\).

\[
\frac{1}{N} \sum_{t=1}^{N} (\ell^q_t[t] - \ell^o_t[t])^2 \leq \frac{n}{c_u^2 \gamma_g^4} \|L^q - L^o\|_F^2 \leq \frac{n}{c_u^2 \gamma_g^4}
\]

This converges to zero at the rate of \(O(N^{-1/2})\) with \(\gamma_g = \Omega(N^{1/8})\).

With properly setting \(\gamma_g\), e.g., \(\gamma_g = \Omega(N^{1/8})\), we can have:

\[
\frac{1}{N} \sum_{t=1}^{N} (\ell^q_t[t] - y_t)^2 = O\left(N^{-1/2}\right)
\]

What does that mean?
SSL with Graphs: What is behind it?

Why and when it helps?

Can we guarantee benefit of SSL over SL?

Are there cases when manifold SSL is provably helpful?

Say $\mathcal{X}$ is supported on manifold $\mathcal{M}$. Compare two cases:

- SL: does not know about $\mathcal{M}$ and only knows $(x_i, y_i)$
- SSL: perfect knowledge of $\mathcal{M} \equiv$ humongous amounts of $x_i$

SSL with Graphs: What is behind it?

Set of learning problems - collections $\mathcal{P}$ of probability distributions:

$$\mathcal{P} = \bigcup_{\mathcal{M}} \mathcal{P}_\mathcal{M} = \bigcup_{\mathcal{M}} \{ p \in \mathcal{P} | p_X \text{ is uniform on } \mathcal{M} \}$$
SSL with Graphs: What is behind it?

Set of problems \( \mathcal{P} = \bigcup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} = \{ p \in \mathcal{P} | p_X \text{ is uniform on } \mathcal{M} \} \)

Regression function \( m_p = \mathbb{E} [y | x] \) when \( x \in \mathcal{M} \)

Algorithm \( A \) and labeled examples \( \bar{z} = \{ z_i \}_{i=1}^{n_l} = \{ (x_i, y_i) \}_{i=1}^{n_l} \)

Minimax rate

\[
R(n_l, \mathcal{P}) = \inf_A \sup_{p \in \mathcal{P}} \mathbb{E}_{\bar{z}} \left[ \| A(\bar{z}) - m_p \|_{L^2(p_X)} \right]
\]

Since \( \mathcal{P} = \bigcup_{\mathcal{M}} \mathcal{P}_{\mathcal{M}} \)

\[
R(n_l, \mathcal{P}) = \inf_A \sup_{\mathcal{M}} \sup_{p \in \mathcal{P}_{\mathcal{M}}} \mathbb{E}_{\bar{z}} \left[ \| A(\bar{z}) - m_p \|_{L^2(p_X)} \right]
\]

(SSL) When \( A \) is allowed to know \( \mathcal{M} \)

\[
Q(n_l, \mathcal{P}) = \sup_{\mathcal{M}} \inf_A \sup_{p \in \mathcal{P}_{\mathcal{M}}} \mathbb{E}_{\bar{z}} \left[ \| A(\bar{z}) - m_p \|_{L^2(p_X)} \right]
\]

In which cases there is a gap between \( Q(n_l, \mathcal{P}) \) and \( R(n_l, \mathcal{P}) \)?
SSL with Graphs: What is behind it?

Hypothesis space $\mathcal{H}$: half of the circle as $+1$ and the rest as $-1$

Case 1: $\mathcal{M}$ is known to the learner ($\mathcal{H}_\mathcal{M}$)

What is a VC dimension of $\mathcal{H}_\mathcal{M}$?

Optimal rate $Q(n, \mathcal{P}) \leq 2\sqrt{\frac{3 \log n_l}{n_l}}$
SSL with Graphs: What is behind it?

**Case 2:** $\mathcal{M}$ is unknown to the learner

$$R(n_l, \mathcal{P}) = \inf_A \sup_{\mathcal{P}} \mathbb{E}_z \left[ \| A(z) - m_p \|_{L^2(p_X)} \right] = \Omega(1)$$

We consider $2^d$ manifolds of the form

$$\mathcal{M} = \text{Loops} \cup \text{Links} \cup C$$

where

$$C = \bigcup_{i=1}^d C_i$$

**Main idea:** $d$ segments in $C$, $d - l$ with no data, $2^l$ possible choices for labels, which helps us to lower bound $\| A(z) - m_p \|_{L^2(p_X)}$
SSL with Graphs: What is behind it?

Knowing the manifold helps

- $C_1$ and $C_4$ are close
- $C_1$ and $C_3$ are far
- we also need: target function varies smoothly
- altogether: closeness on manifold $\rightarrow$ similarity in labels
SSL with Graphs: What is behind it?

What does it mean to know $\mathcal{M}$?

Different degrees of knowing $\mathcal{M}$

- set membership oracle: $x \in \mathcal{M}$
- approximate oracle
- knowing the harmonic functions on $\mathcal{M}$
- knowing the Laplacian $\mathcal{L}_\mathcal{M}$
- knowing eigenvalues and eigenfunctions
- topological invariants, e.g., dimension
- metric information: geodesic distance
Scaling SSL with Graphs to Millions

Semi-supervised learning with graphs

\[ f^* = \min_{f \in \mathbb{R}^N} (f - y)^T C (f - y) + f^T L f \]

Let us see the same in eigenbasis of \( L = U \Lambda U^T \), i.e., \( f = U \alpha \)

\[ \alpha^* = \min_{\alpha \in \mathbb{R}^N} (U \alpha - y)^T C (U \alpha - y) + \alpha^T \Lambda \alpha \]

What is the problem with scalability?

Diagonalization of \( N \times N \) matrix

What can we do? Let’s take only first \( k \) eigenvectors \( f = U \alpha \! \)
Scaling SSL with Graphs to Millions

\[ \mathbf{U} \text{ is now a } n \times k \text{ matrix} \]

\[ \alpha^* = \min_{\alpha \in \mathbb{R}^n} (\mathbf{U}\alpha - \mathbf{y})^\top \mathbf{C}(\mathbf{U}\alpha - \mathbf{y}) + \alpha^\top \Lambda \alpha \]

Closed form solution is \((\Lambda + \mathbf{U}^\top \mathbf{C}\mathbf{U})\alpha = \mathbf{U}^\top \mathbf{C}\mathbf{y}\)

What is the size of this system of equation now?

Cool! Any problem with this approach?

Are there any reasonable assumptions when this is feasible?

Let’s see what happens when \(N \rightarrow \infty\)!
Scaling SSL with Graphs to Millions

Density

Data

Landmarks

Limit as $n \to \infty$

Linear in number of data-points

Polynomial in number of landmarks

https://cs.nyu.edu/~fergus/papers/fwt_ssl.pdf
Scaling SSL with Graphs to Millions

What happens to $L$ when $N \to \infty$?

We have data $x_i \in \mathbb{R}$ sampled from $p(x)$.

When $n \to \infty$, instead of vectors $f$, we consider functions $F(x)$.

Instead of $L$, we define $L_p$ - **weighted smoothness operator**

$$L_p (F) = \frac{1}{2} \int (F(x_1) - F(x_2))^2 W(x_1, x_2) p(x_1) p(x_2) \, dx_1 dx_2$$

with $W(x_1, x_2) = \frac{\exp(-\|x_1-x_2\|^2)}{2\sigma^2}$

$L$ defined the eigenvectors of increasing smoothness.

**What defines $L_p$?** Eigenfunctions!
Scaling SSL with Graphs to Millions

$$\mathcal{L}_p(F) = \frac{1}{2} \int (F(x_1) - F(x_2))^2 W(x_1, x_2)p(x_1)p(x_2) dx_1 dx_2$$

First eigenfunction

$$\Phi_1 = \arg \min_{F: \int F^2(x)p(x)D(x) \, dx = 1} \mathcal{L}_p(F)$$

where $$D(x) = \int_{x_2} W(x, x_2)p(x_2) \, dx_2$$

What is the solution? $$\Phi_1(x) = 1$$ because $$\mathcal{L}_p(1) = 0$$

How to define $$\Phi_2$$? Same, constraining to be orthogonal to $$\Phi_1$$

$$\int F(x) \Phi_1(x) p(x) D(x) \, dx = 0$$
Scaling SSL with Graphs to Millions

Eigenfunctions of $\mathcal{L}_p$

$\Phi_3$ as before, orthogonal to $\Phi_1$ and $\Phi_2$ etc.

How to define eigenvalues? $\lambda_k = \mathcal{L}_p (\Phi_k)$

Relationship to the discrete Laplacian

$$\frac{1}{N^2} f^T L f = \frac{1}{2N^2} \sum_{ij} W_{ij} (f_i - f_j)^2 \xrightarrow{N \to \infty} \mathcal{L}_p (F)$$

http://www.informatik.uni-hamburg.de/ML/contents/people/luxburg/publications/Luxburg04_diss.pdf

Isn’t estimating eigenfunctions $p(x)$ more difficult?

Are there some “easy” distributions?

Can we compute it numerically?
Scaling SSL with Graphs to Millions

**Eigenvectors**

Data |
- |
- |

$\phi_1, \sigma_1 = 0$

$\phi_2, \sigma_2 = 0.0002$

$\phi_3, \sigma_3 = 0.038$

**Eigenfunctions**

Density |
- |
- |

$\Phi_1, \sigma_1 = 0$

$\Phi_2, \sigma_2 = 0.0002$

$\Phi_3, \sigma_3 = 0.035$
Scaling SSL with Graphs to Millions

Factorized data distribution

What if

\[ p(s) = p(s_1) p(s_2) \ldots p(s_d) \]

In general, this is not true. But we can rotate data with \( s = Rx \).

Treating each factor individually

\[ p_k \overset{\text{def}}{=} \text{marginal distribution of } s_k \]

\[ \Phi_i(s_k) \overset{\text{def}}{=} \text{eigenfunction of } \mathcal{L}_{p_k} \text{ with eigenvalue } \lambda_i \]

Then: \( \Phi_i(s) = \Phi_i(s_k) \) is eigenfunction of \( \mathcal{L}_p \) with \( \lambda_i \)

We only considered single-coordinate eigenfunctions.
How to approximate 1D density? Histograms!

Algorithm of Fergus et al. [FWT09] for eigenfunctions

- Find $\mathbf{R}$ such that $\mathbf{s} = \mathbf{Rx}$
- For each “independent” $s_k$ approximate $p(s_k)$
- Given $p(s_k)$ numerically solve for eigensystem of $\mathcal{L}_{p_k}$

$$
\left(\widehat{\mathbf{D}} - \mathbf{P}\overline{\mathbf{W}}\mathbf{P}\right) \mathbf{g} = \lambda \mathbf{P}\overline{\mathbf{D}}\mathbf{g} \quad \text{(generalized eigensystem)}
$$

- $\mathbf{g}$ - vector of length $B \equiv$ number of bins
- $\mathbf{P}$ - density at discrete points
- $\widehat{\mathbf{D}}$ - diagonal sum of $\mathbf{P}\overline{\mathbf{W}}\mathbf{P}$
- $\overline{\mathbf{D}}$ - diagonal sum of $\mathbf{P}\overline{\mathbf{W}}$
- Order eigenfunctions by increasing eigenvalues

https://cs.nyu.edu/~fergus/papers/fwt_ssl.pdf
Scaling SSL with Graphs to Millions

Numerical 1D Eigenfunctions

1\textsuperscript{st} Eigenfunction of $h(x_1)$

2\textsuperscript{nd} Eigenfunction of $h(x_1)$

3\textsuperscript{rd} Eigenfunction of $h(x_1)$

https://cs.nyu.edu/~fergus/papers/fwt_ssl.pdf
Scaling SSL with Graphs to Millions

Computational complexity for $N \times d$ dataset

**Typical harmonic approach**

one diagonalization of $N \times N$ system

**Numerical eigenfunctions** with $B$ bins and $k$ eigenvectors

$d$ eigenvector problems of $B \times B$

\[
\left( \tilde{D} - P\tilde{W}P \right) g = \lambda P\tilde{D}g
\]

one $k \times k$ least squares problem

\[
(\Lambda + U^TCU)\alpha = U^TCy
\]

some details: several approximation, eigenvectors only linear combinations single-coordinate eigenvectors, . . .

**When $d$ is not too big then $N$ can be in millions!**
Scaling SSL with Graphs to Millions

CIFAR experiments https://cs.nyu.edu/~fergus/papers/fwt_ssl.pdf