Beyond Supervised Classification: Extreme Minimal Supervision with the Graph 1-Laplacian

Angelica I. Aviles-Rivero
DPMMS, University of Cambridge,
Wilforce Road, UK,
ai323@cam.ac.uk

Nicolas Papadakis
IMB, Université Bordeaux,
33405 Talence Cedex, France
nicolas.papadakis@
math.u-bordeaux.fr

Ruoteng Li
NUS,
Singapore,
liruoteng@u.nus.edu

Samar M Alsaleh
GWU,
2121 I St NW, USA,
sm57@gwu.edu

Robby T Tan
Yale-NUS College,
Singapore,
robby.tan@yale-nus.edu.sg

Carola-Bibiane Schönlieb
DAMPT, University of Cambridge,
Wilforce Road, UK,
cbs31@cam.ac.uk

Abstract

We consider the task of classifying when an extremely reduced amount of labelled data is available. This problem is of a great interest, in several real-world problems, as obtaining large amounts of labelled data is expensive and time consuming. We present a novel semi-supervised framework for multi-class classification that is based on the non-smooth ℓ_1 norm of a normalised Dirichlet energy based on the graph Laplacian. Our transductive framework is framed under a novel functional with carefully selected class priors – that enforces a sufficiently smooth solution that strengthens the intrinsic relation between the labelled and unlabelled data. We demonstrate through extensive experimental results on large datasets CIFAR-10 and ChestX-ray14, that our method outperforms classic methods and readily competes with recent deep-learning approaches.

1 Introduction

In this era of big data, deep learning (DL) has reported astonishing results for different tasks in computer vision including image classification e.g. [21, 18], detection and segmentation just to name few. In particular, for the task of image classification, a major breakthrough has been reported in the setting of supervised learning. In this context, majority of methods are based on deep convolutional neural networks including ResNet [16], VGG [27] and SE-Net [18] in which pre-trained, fine tuned and trained from scratch solutions have been considered. A key factor, for these impressive results, is the assumption of a large corpus of labelled data. These labels can be generated either by humans or automatically on proxy tasks. However, to obtain well-annotated labels is expensive and time consuming, and one should account for either human bias and uncertainty that adversely effect the classification output. These drawbacks have motivated semi-supervised learning (SSL) to be a focus of great interest in the community.

Labelling data is helpful but it is expensive and difficult to do

What is the difference between semi-supervised and unsupervised?

SSL is when there is some labelled data, but not lots of it. So you have a mixture of labelled and unlabelled data.

The key idea of SSL is to exploit both labelled and unlabelled data to produce a good classification output. The desirable advantages of this setting is that one decreases the dependency for a large amounts of well-annotated data whilst gaining further understanding of the relationships in the data. A comprehensive revision on SSL can be seen in [9]. In the transductive setting, several algorithmic approaches have been proposed such as [37, 35, 30, 36, 20, 34] whilst in the inductive setting also promising results have been reported including [32, 28]. More recently, DL for semi-supervised

Preprint.

learning has been explored in both settings such as in [22, 28, 19]. We refer the reader to [14, 10] for a detailed revision on SSL for image classification.

In this work, we focus on the transductive setting for image classification with the normalised Dirichlet energy (1) based on the graph Laplacian. Although promising results have been shown in this context, for example, the seminal algorithm of [35] was introduced to perform such a graph transduction through the propagation of few labels by the minimisation of energy (1) for $p = 2$. Latter machine learning studies nevertheless showed that the use of non-smooth energies with the $p = 1$ norm, related to non local total variation, can achieve better clustering performances [7], but original algorithms were only approximating $p \to 1$.

More advanced optimisation tools were therefore proposed to consider the exact $p = 1$ norm for binary [17] or multi-class [4] graph transduction. As underlined in [29], the normalisation of the operator is nevertheless crucial, to ensure within-cluster similarity when the degrees of the nodes d_i are broadly distributed in the graph.

Contributions. In order to address these different issues, we propose a new graph based semi-supervised framework called EMS-1L. The novelty of our framework largely relies on:

- A new multi-class classification functional based on the normalised and non-smooth ($p = 1$) energy (1), where the selection of carefully chosen class priors enforces a sufficiently smooth solution that strengthens the intrinsic relation between the labelled and unlabelled data.
- We demonstrate that our framework accurately learns to classify different challenging datasets such as ChestX-ray14, with a performance comparable to state of the art DL techniques, whilst using an extremely smaller amount of labelled data.
- We show that our framework can be extended to deep SSL, and that it achieves the lowest error rate in comparison with state-of-the-art SSL approaches on CIFAR-10 dataset.

A metric of the usefulness of the algorithm here is its ability to use unlabelled data. The paper is trying to find an algorithm that reduces dependency on labelled data whilst keeping a good performance. This is important, we can't just base the validity of the algorithm on its performance but also on the amount of labelled data it requires.

2 Extreme Minimal Supervision with the Normalised Dirichlet $1-$Energy: Preliminaries

So our problem context is that we have LOTS of unlabelled data but not much labelled. e.g for an ancient manuscript language this might be the case.

Formally speaking, we aim at solving the following problem. Given a small amount of labeled data $\{(x_i, y_i)\}_{i=1}^{l}$ with provided labels $\mathcal{L} = \{1, .., L\}$ and $\{y_i\}_{i=1}^{l} \in \mathcal{L}$ and a large amount of unlabelled data $\{x_k\}_{k=l+1}^{n}$, we seek to infer a function $f : \mathcal{X}^n \mapsto \mathcal{Y}^n$ such that f gets a good estimate for $\{x_k\}_{k=l+1}^{l+n}$. This problem is illustrated in Figure 1, where visualisations were obtained from one of our experiments. For addressing this problem, we consider functions $u \in \mathbb{R}^n$ defined over a set $\mathcal{N}$ of n nodes. The main focus of interest in this work are convex and absolutely p-homogeneous (i.e. $J(\alpha u) = |\alpha|^p J(u)$) non-local functionals of the form:

$$D_p(u) = \sum_{ij} w_{ij} \left\| \frac{u_i}{d_i^{1/p}} - \frac{u_j}{d_j^{1/p}} \right\|^p , \tag{1}$$

Weights are symmetric d is a vector of row sums of the weights. Why is there normalisation by this d_i?

Why is d also the vector of degrees if the nodes? Is this a constraint on the weights?

with weights $w_{ij} = w_{ji} \geq 0$ taken such that the vector $d \in \mathbb{R}^n$ has non null entries satisfying: $d_i = \sum_j w_{ij} > 0$. This energy acts on the graph defined by nodes $\mathcal{N}$ and weights w. With respect to classical Dirichlet energies associated to the graph p-Laplacian [1, 12, 17, 4], it includes a normalisation through the rescaling with the degree of the node. We will focus our attention to the non smooth case $p = 1$ with the absolutely one homogeneous energy defined by the function $J(u) = D_1(u)$ that can be rewritten as:

$$J(u) = \|WD^{-1}u\|_1, \tag{2}$$

with a $n \times n$ diagonal matrix $D = \text{diag}(d)$, containing the nodes degree so that $d = D\mathbf{1}_n$, and a $m \times n$ matrix W that encodes the m edges in the graph. Each of these edges is represented on a different line of the sparse matrix W, with the value w_{ij} (resp. $-w_{ij}$) on the column i (resp. j).

Note that W is very sparse and has row sum zero by definition. Each row represents a graph edge and the non-zero elements of a row represent the two nodes connected to eachother. The graph is undirected since the weights were defined as symmetric (! This would not be the case for much social media interaction!)

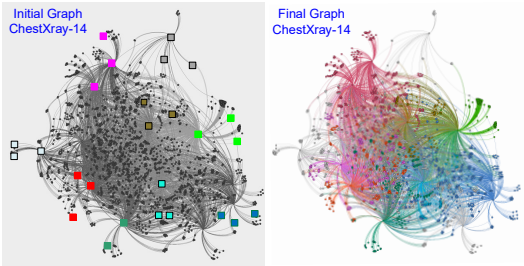

Figure 1: Graphical representation of one our experiments, where in the final classified graph, each colour represents a different class

Look on Wikipedia to understand more about what a sub-differential is. In this case, I believe it is necessary since we are using norm 1 which is not differentiable so instead of the nice theory of derivatives we have subdifferentials.

This section is not understood, an exposition of this in an essay would need much more clarification and definitions.

Subdifferential Let us first define as ∂J the set of possible subdifferentials of J: $\partial J = \{p, \text{ s.t. } \exists u, \text{ with } p \in \partial J(u)\}$. Any absolutely one homogeneous function J checks:

$$J(u) = \sup_{p \in \partial J} \langle p, u \rangle \tag{3}$$

so that $J(u) = \langle p, u \rangle$, $\forall p \in \partial J(u)$.

For the particular function J defined in (2), we can observe that

$$p \in \partial J \Leftrightarrow p = D^{-1}W^\top z, \text{ with } ||z||_\infty \leq 1. \tag{4}$$

Considering the finite dimension setting, there exists $L_J < \infty$ such that $\|p\|_2 < L_J, \forall p \in \partial J$. We also have the following property.

Proposition 1. *For all $p \in \partial J$, with J defined in* (2)*, one has*

$$\langle p, d \rangle = 0.$$

Proof. Observing that $d = D\mathbf{1}$ and using (4) we have that $\exists z \in \mathbb{R}^m$ such that

$$\langle p, d \rangle = \langle D^{-1}W^\top z, D\mathbf{1}_n \rangle.$$

Since the weights satisfy $w_{ij} = w_{ji}$, then for all $z \in \mathbb{R}^m$:

$$\langle W^\top z, \mathbf{1}_n \rangle = \sum_i \sum_j w_{ij}(z_i - z_j) = \sum_i \sum_{j>i} w_{ij}(z_i - z_j - z_i + z_j) = 0.$$

□

Eigenfunction. Eigenfunctions of any functional J satisfy $\lambda u \in \partial J(u)$. For J being the nonlocal total variation, (i.e. when d_i is constant), eigenfunctions are known to be essential tools to provide a relevant clustering of the graph [29]. Methods [7, 3, 5, 2, 13] have thus been designed to estimate such eigenfunctions through the local minimisation of the Rayleigh quotient, which reads:

$$\min_{\|u\|_2=1} \frac{J(u)}{H(u)}, \tag{5}$$

with another absolutely one homogeneous function H, that is typically a norm. Taking $H(u) = \|u\|_2$ as the ℓ_2 norm, one can recover eigenfunctions of J [2]. For $H(u) = \|u\|_1$ being the ℓ_1 norm, these approaches can compute bi-valued functions u that are local minima of (5) and eigenfunctions of J [13]. Being bivalued, these estimations can easily be used to realise a partition of the domain. Such schemes also relate to the Cheeger cut of the graph induced by nodes u_i and edges w_{ij}. Balanced cuts can also be obtained by considering $H(u) = \|u - \text{median}(u)\|_1$ [4, 13].

A last point to underline comes from Proposition 1, that states that eigenfunctions $\lambda u \in \partial J(u)$ should be orthogonal to d. It is thus important to design schemes that ensure this property.

3 Classifying under Extreme Minimal Supervision the Normalised Dirichlet $1-$Energy

In the following, instead of u_i, we will denote by $u(x)$ the value of function u at node x. In order to realise a binary partition of the domain of the graph $\mathcal{N}$ through the minimisation of the quotient $R(u) = J(u)/H(u)$, we adapt the method of [13] to incorporate the scaling $d(x)$ of (2) and consider the semi-explicit PDE:

$$\begin{cases} \frac{u_{k+1/2}-u_k}{\delta t} & = \frac{J(u_k)}{H(u_k)}(q_k - \tilde{q}_k) - p_{k+1/2}, \\ u_{k+1} & = \frac{u_{k+1/2}}{\|u_{k+1/2}\|_2} \end{cases} \tag{6}$$

with $p_{k+1/2} \in \partial J(u_{k+1/2})$, $q \in \partial H(u_k)$, $\tilde{q}_k = \frac{\langle d, q_k \rangle}{\langle d, d \rangle} d$. We recall that both J and H are absolutely one homogeneous and satisfy (3). Since $\langle p, d \rangle = 0$, $\forall p \in \partial J$, the shift with $\tilde{q}_k$ is necessary to show the convergence of the PDE as we have $u_k \to u^* \Rightarrow \frac{J(u^*)}{H(u^*)}(q^* - \tilde{q}^*) = p^*$, for $p^* \in \partial J(u^*)$ and $q^* \in \partial H(u^*)$.

Such sequence u_k satisfies the following properties.

Proposition 2. *For $\langle u_0, d\rangle = 0$, the trajectory u_k given by* (6) *satisfies*

1 $\langle u_{k+1}, d\rangle = 0$,

2 $||u_{k+1/2}||_2 \geq ||u_k||_2$,

3 $R(u_k)$ *is non increasing,*

4 $H(u_{k+1/2}) \leq \kappa < +\infty$.

The proof is given in the Supplementary Material. It namely uses the fact that $u_{k+1/2}$ is the unique minimiser of:

$$F_k(u) = \frac{1}{2\delta t}\|u - u_k\|_2^2 + R(u_k)\langle q_k - \tilde{q}_k, u\rangle + J(u). \tag{7}$$

Hence, we can show the convergence of the trajectory.

Proposition 3. *The sequence u_k defined in* (6) *converges to a non constant steady point u^*.*

Proof. As $u_{k+1/2}$ is the unique minimizer of F_k in (7) that checks $F_k(u_k) = 0$, and as we have $\langle q_k - \tilde{q}_k, u_{k+1/2}\rangle \leq H(u_{k+1/2})$, we get

$$\frac{1}{2\delta t H(u_{k+1/2})}\|u_{k+1/2} - u_k\|_2^2 + R(u_{k+1}) \leq R(u_k), \tag{8}$$

Since u_{k+1} is the orthogonal projection of $u_{k+1/2}$ on the ℓ_2 ball then $\|u_{k+1} - u_k\|_2^2 \leq \|u_{k+1/2} - u_k\|_2$. Finally, from point 4 of Proposition 2, we have that $1/H(u_{k+1/2}) \geq 1/\kappa$. We then sum relation (8) from 0 to K and deduce that:

$$\sum_{k=0}^{K} \frac{1}{2\delta t \kappa}\|u_{k+1} - u_k\|_2^2 \leq H(u_0).$$

so that $\|u_{k+1} - u_k\|_2$ converges to 0. Since all the quantities are bounded, we can show (see [13], Theorem 2.1) that up to a subsequence $u_k \to u^*$.

From Proposition 2, the points u_k being of constant norm and $\langle d, u_k\rangle$ being zero (with positive weights d_i), the limit point u^* of the trajectory (6) necessarily has negative and positive entries. □

In practice, to realise a partition of the graph with the scheme (6), we miniminise the functional (7) at each iteration k with the primal dual algorithm in [8] to obtain $u_{k+1/2}$, and then normalise this estimation. As it is non constant and satisfies $\langle u^*, d\rangle = 0$, the limit of the scheme u^* can be used for partitioning with the simple criteria $u^* > 0$.

Multi-class clustering. We now aim at finding L coupled functions u^l that are all local minima of the ratio $J(u)/H(u)$. The issue is to define a good coupling constraint between the u^l's such that it is easy to project on. Let $\mathbf{u} = [u^1, \cdots u^L]$, we here consider the simple linear coupling :

$$C : \{\mathbf{u}, \text{ s.t. } \sum_{l=1}^{L} u^l(x) = 0, \, \forall x \in \mathcal{N}\}. \tag{9}$$

There are three main reasons for considering such coupling instead of classical simplex [4, 24, 15] or orthogonality [11] constraints:

1 Projection on this linear constraint is explicit with a simple shift of the vector $\mathbf{u}(x)$ for each node x. On the other hand, simplex constraint ($u^l(x) \geq 0$, $\sum_l u^l(x) = 1$, $\forall x$) requires more expensive projections of the vectors $\mathbf{u}(x)$ on the L simplex. Last, projection on the orthogonal constraint of the u^l's is a non convex problem.

2 Contrary to the simplex constraint, it is compatible with the weighted zero mean condition $\langle u^l, d\rangle$ that any eigenfunction of J should satisfy, as shown in Proposition 1.

3 The characteristic function of a linear constraint is absolutely one homogeneous. This leads to a natural extension of the binary case.

Multi-class flow. We now consider the problem:

$$\min_{\|\mathbf{u}\|_2=1} \sum_{l=1}^{L} \frac{J(u^l)}{H(u^l)}. \tag{10}$$

To find a local minima of (10), we define the iterative multi-class functional, which reads:

$$F_k^L(\mathbf{u}) = \frac{1}{2\delta t}\|\mathbf{u}-\mathbf{u}_k\|_2^2 - \sum_{l=1}^{L} R(u_k^l)\langle q_k^l - \tilde{q}_k^l, u^l\rangle + \sum_{l=1}^{L} J(u^l) + \chi_C(\mathbf{u}) \tag{11}$$

where $q_k^l \in \partial H(u_k^l)$ and χ_C is the characterstic function of the constraints (9). Starting from an initial point $\mathbf{u}_0$ that satisfies the constraint ($\chi_C(\mathbf{u}_0)=0$) and has been normalised ($\|\mathbf{u}_0\|_2^2 = \sum_{l=1}^{L}\|u_0^l\|_2^2 = 1$), the scheme we consider reads:

$$\left\{ \begin{array}{ll} u_{k+1/2}^l & = u_k^l + \delta t\left(R(u_k^l)(q_k^l - \tilde{q}_k^l) - p_{k+1/2}^l - r_{k+1/2}^l\right) \\ \mathbf{u}_{k+1} & = \frac{\mathbf{u}_{k+1/2}}{\|\mathbf{u}_{k+1/2}\|_2} \end{array} \right. \tag{12}$$

where $p_{k+1/2}^l \in \partial J(u_{k+1/2}^l)$ and $\mathbf{r}_{k+1/2} \in \partial\chi_C(\mathbf{u}_{k+1/2})$, and the point $\mathbf{u}_{k+1/2}$ in the above PDE corresponds to the global minimiser of (11). Notice that the subgradient of the one homogeneous functional χ_C can be characterised with:

$$\mathbf{r} \in \partial\chi_C \Rightarrow \{r^l(x) = \alpha(x),\, \forall l = 1\cdots L \text{ and } x \in \mathcal{N}\}. \tag{13}$$

In practice, if for some l, $u_{k+1/2}^l$ vanishes, then we define $R(u_{k+1}^l) = 0$ for the next iteration. With such assumptions, the sequence $\mathbf{u}_k$ have the following properties, that are shown in Supplementary Material.

Proposition 4. *For $\langle u_0^l, d\rangle = 0$, $l = 1\cdots L$, the trajectory $\mathbf{u}_k$ given by* (12) *satisfies*

1 $\langle u_k^l, d\rangle = 0$,

2 $||\mathbf{u}_k||_2 \leq ||\mathbf{u}_{k+1/2}||_2 \leq \kappa < \infty$,

3 $\sum_{l=1}^{L} H(u_{k+1}^l)\left(R(u_{k+1}^l) - R(u_k^l)\right) \leq -\frac{1}{2\delta t\kappa}\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2$.

Point 3 of Proposition 4 contains weights $H(u_{k+1}^l)$ that prevent from showing the exact decrease of the sum of ratios. This is thus similar to the approach in [4].

To ensure the decrease of the sum of ratios $\sum_{l=1}^{L} J(u_k^l)/H(u_k^l)$, is is possible to introduce auxiliary variables dealing with individual ratio decrease, as in [24]. The involved sub-problem at each iteration k is nevertheless more complex to solve.

Also notice that as there is no prior information on nodes' labels, clusters can vanish or 2 clusters may become proportional one to the other. Such issues can nevertheless not happen in the transductive setting we now consider.

Label Propagation: Multi-Class Classification. The previous settings are unsupervised. We now consider a semi-supervised setting where we know small subsets of labeled nodes $\mathcal{N}^l \subset \mathcal{N}$ (with $|\mathcal{N}^l| << | \subset \mathcal{N}|$) belonging to each cluster i, with $\mathcal{N}^l \cap \mathcal{N}_j = \emptyset$. Denoting $\mathcal{L} = \cup_{l=1}^{L}\mathcal{N}^l$, the objective is to propagate the prior information in the graph in order to predict the labels of the remaining nodes $x \in \mathcal{N}\backslash\mathcal{L}$. To that end, we simply have to modify the coupling constraint C in (9) as

$$C : \left\{ \begin{array}{ll} \sum_{l=1}^{L} u^l(x) = 0 & \text{if } x \in \mathcal{N}\backslash\mathcal{L} \\ \mathbf{u},\ \text{s.t.} \quad u^l(x) \geq \epsilon & \text{if } x \in \mathcal{N}^l \\ u^{l'}(x) \leq -\epsilon, \forall l' \neq l & \text{if } x \in \mathcal{L}\backslash\mathcal{N}^l \end{array} \right\}. \tag{14}$$

With such constraint, clusters can no more vanish or merge since they all contain different active nodes $x \in \mathcal{N}^l$ satisfying $u^l(x) > 0$. The same PDE (12) can be applied to propagate these labels. Once it has converged, the label of each node $x \in \mathcal{N}\backslash\mathcal{L}$ is taken as:

$$L(x) \in \underset{i\in\{1,\cdots L\}}{\text{argmax}}\ u^l(x).$$

Soft labelling can either be obtained by considering all the clusters with non negative weights $\mathcal{I}(x) = \{l,\, u^l(x) \geq 0\} \neq \emptyset$ with relative weights $w^l(x) = u^l(x)/(\sum_{l\in\mathcal{I}(x)} u^l(x))$ and the convention that $w^l(x) = 1/L$, in the case (that has never been observed in our experiments) that $u^l(x) = 0$ for all $l = 1 \cdots L$.

The parameter ϵ in (14) is set to a small numerical value. Indeed, even if $\mathbf{u}_{k+1/2} \in C$ by construction, a small ϵ is required to ensure that, after the rescaling, $\mathbf{u}_{k+1} = \mathbf{u}_{k+1/2}/\|\mathbf{u}_{k+1/2}\|_2 \in C$. One can consider different values ϵ^l for each class. In the case where $L = 2$, d is constant and $H(u) = \|u - \mathrm{median}(u)\|_1$, u^l is expected to be bivalued [13] and the value of ϵ has a clear meaning. In that framework, $\epsilon = 1/\sqrt{|\mathcal{N}|(|\mathcal{N}|-1)}$ corresponds to no prior on the size of the clusters, whereas $\epsilon = 1/\sqrt{|\mathcal{N}|n}$ encourage the clusters to be of homogeneous size.

4 Experimental Results

This section is focused on describing in detail the experiments that we conducted to evaluate our proposed approach.

4.1 Implementation Details

We here describe the specifics of our experimental setting including the data description and the evaluation methodology.

Data Description. We validate our approach using three datasets - one small-scale and two large-scale datasets. 1) UCI ML hand-written digits dataset, we use the test set composed of 1797 images of size 8×8, and 10 classes. We also use 2) ChestX-ray14 dataset [31], which is composed of 112,120 frontal chest view X-ray with size of 1024×1024. The dataset is composed of 14 classes. 3) The CIFAR-10 datase contains 60,000 color images of size 32×32 and 10 different classes. All classification results were performed using these datasets.

Evaluation Protocol. We design the following evaluation scheme to validate our theory. Firstly, we evaluate our proposed EMS-1L approach against two classic methods: Label Propagation (LP) [37] and Local to global consistency (LCG) [35]. For output quality evaluation, we computed the error rate and F1-score. Secondly and using ChestX-ray14 dataset [31], we compared our approach against two deep learning approaches - WANG17[31] and YAO18 [33]. The quality of the classification was performed by a ROC analysis using the area under the curve (AUC). Finally, we demonstrate that our method can be extended to deep SSL, which evaluation is performed on the CIFAR-10 dataset and compared against state-of-the-art deep SSL[25, 26, 28, 19] and a fully supervised technique [23]. For this part, we evaluate the quality of the classifiers by reporting the error rate for a range of number of labelled samples.

Each experiment has been repeated 10 time and the average and standard deviation are reported. For the compared methods, the parameters were set using the default values provided in the demo code or referenced in the papers themselves.

4.2 How good is EMS-1L?

We start by giving some insight into the performance of our approach with a comparison against two classic methods LP [37] and LCG [35], which results, using the digits dataset, are reported in Table 1. One can see that for all metrics and percentages of labeled samples, our approach outperforms the compared methods by a significant margin. In particular, one can observe that with even 1% of labelled data, the error rate of our EMS-1L approach is almost half the second best method which is extrapolated to the remaining percentages of labeled samples and evaluation metrics. This shows that our EMS-1L approach is outperforms the compared methods even under extremely minimal labeled samples.

To further evaluate the results of our approach, we move to a large scale dataset ChestX-ray14. Our motivation to use this dataset is coming from a central problem in medical imaging which is the lack of reliable quality annotated data. In particular, the interpretation of X-ray data heavily relies on the radiologist's expertise and there is still a substantial clinical error on the outcome [6]. We ran our

Metric	Method	Percentage of Labeled Samples				
		1%	2%	5%	10%	20%
Error Rate	LP [37]	40.53±5.38	28.91±4.01	22.70±3.23	10.04±1.49	5.83±1.38
	LCG [35]	29.57±8.22	11.00±3.09	9.63±2.41	5.16±1.45	3.44±1.28
	EMS-1L	**14.21±5.63**	**6.51±1.86**	**3.46±0.91**	**1.80±0.54**	**1.09±0.24**
F1-Micro	LP [37]	59.48±6.99	67.66±5.15	76.08±3.67	89.86±1.53	94.12±1.46
	LCG [35]	63.80±10.74	88.23±4.16	89.95±2.89	94.80±1.49	95.55±1.28
	EMS-1L	**84.50±7.48**	**93.40±1.98**	**96.52±0.93**	**98.20±0.11**	**98.91±0.24**
F1-Macro	LP [37]	56.48±5.38	71.09±4.01	77.30±3.22	89.96±1.49	94.17±1.38
	LCG [35]	70.43±8.22	89.00±3.09	90.37±2.41	94.84±1.45	95.56±1.28
	EMS-1L	**85.79±5.63**	**93.49±1.86**	**96.54±0.91**	**98.54±0.91**	**98.91±0.24**

Table 1: Compression with state of the art classic transductive methods on the Digits dataset

Approach	Average AUC
WANG17[31]	0.7451
YAO18 [33]	0.7614
MT [28]	0.5
EMS-1L (20%)	0.7888

Table 2: Comparison of the classification accuracy of EMS-1L against three state-of-the-art deep learning method

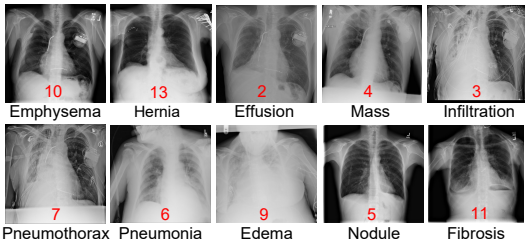

Figure 2: Examples of correct classifications produced by our framework

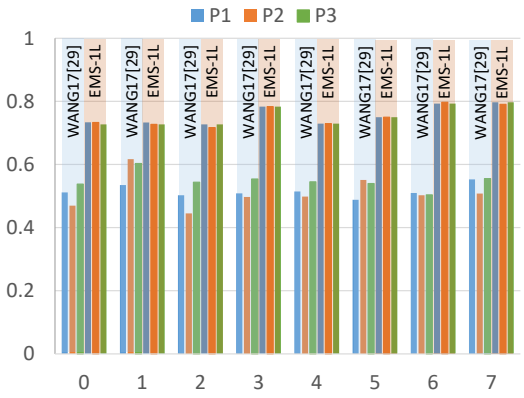

Figure 3: Plot highlighting the sensitivity of the AUC for each class when changing the data partition of the data set (using 15% for training)

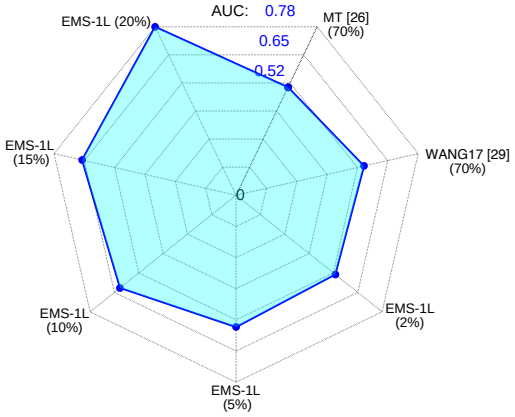

Figure 4: Comparison of the classification accuracy of EMS-1L,using different amounts of labelled data, against state-of-the-art methods.

approach and compared against two state-of-the-art works on X-ray classification WANG17[31] and YAO18 [33], which are supervised methods and, therefore, assume a large corpus of annotated data.

In Figure 2, we show few sample output that were correctly classified by our approach. Table 2 shows the averaged AUC for all classes of our approach compared against WANG17 [31], YAO18 [33], and MT [28] using the official data partition. From a inspection in the table, one can see that our EMS-1L approach outperformed the compared methods with only 20% of the data whilst the compared approaches rely on 70% of annotated data.

Moreover, we noticed that the classification output is very stable with respect to changes in the partition of the dataset, which is due to the semi-supervised nature of our EMS-1L approach. This is well reflected in the Figure 3 where we show the AUC results of both EMS-1L and WANG17 [31] using three different random data partitions, including the partition suggested by WANG17 [31]. The

METHOD	LABELLED SAMPLES		
	1000	2000	4000
SNGT [23] (Fully Supervised)	46.43±1.21	33.94±0.73	20.66±0.57
SSL-GAN [25]	21.83±2.01	19.61±2.09	18.63±2.32
TDCNN [26]†	32.67±1.93	22.99±0.79	16.17±0.37
MT [28]	21.55±1.48	15.73±0.31	12.31±0.28
DSSL [19] (diffusion+W)†	22.02±0.88	15.66±0.35	12.69±0.29
Deep EMS-1L	**20.45±1.08**	**13.91±0.23**	**11.08±0.24**

Table 3: Comparison with state of the art methods on semi-supervised learning and as a base line a fully supervised approach on CIFAR-10 dataset. † indicates scores reported in the corresponding work.

plot shows that WANG17 is sensitive to changes in partition which can be explained by the fact that supervised methods heavily rely on the training set being representative. On the other hand, EMS-1L had minimal change in the performance over the three different partitions as the underlying graphical representation is invariant to the partition.

To further analyse the dependency on the portioning and show the advantage of EMS-1L, we compare the AUC results of EMS-1L against WANG17 and MT17 using a random data partitions. The results are reported in Figure 4 - it shows that EML-1L produces a more accurate classification using only 2% of the data labels than WANG17 or MT17 methods do using 70% of the data labels. The plot also shows that as we feed EML-1L more data labels, the classification accuracy increases and significantly outperforms compared approached whilst still using a far smaller amount of data labels.

4.3 Deep EMS-1L: An Alternative View

One interesting observation about our proposed framework is the fact https://www.overleaf.com/project/5cf5a22625299828a84c2376that it can be adapted to DL for semi-supervised learning SSL. To show this ability, we followed the philosophy of [19] in which they considered the seminal work LCG [35]. We used their pseudo-labelling approach and connected our EMS-1L (i.e. we replace LCG with our approach). Then we performed the image classification task on the CIFAR-10 dataset for different label sample counts.

The results of this experiment can be seen in Table3 in which we show as a baseline a fully supervised approach [23] followed by four state of the art DL semi-supervised approaches [25, 26, 28, 19]. One can observe that lowest error rate across different counts of labelled samples is achieved by our extension Deep EMS-1L. After a detailed inspection of the table, we observe that even though the outputs generated with SSL-GAN [25] started close to our score, they were not significantly improved with the increased number of samples.

5 Conclusion

In this work, we addressed the problem of classifying under minimal supervision (i.e. SSL), in particular, in the transductive setting. We proposed a new semi-supervised framework which is framed under a novel optimisation model for the task of image classification. From extensive experimental results, we found the following. Firstly, we showed that our approach significantly outperformed the classic SSL methods. Secondly, we evaluated our EMS-1L method for the task of X-ray classification and demonstrated that our approach competes against the state-of-the-art results in this context whilst requiring an extremely minimal amount of labelled data. Finally, to demonstrate the capabilities of our approach, we showed that it can be extended as a Deep SSL framework. In this context we observed the lowest error rate results on the CIFAR-10 with respect to the state-of-the-art SSL methods. Future work will include investigation of our approach in terms of data aggregation and how to handle unseen classes.

Acknowledgments This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 777826. Support from the CMIH, University of Cambridge is greatly acknowledged.

References

[1] F. Andreu, J. Mazón, J. Rossi, and J. Toledo. A nonlocal p-laplacian evolution equation with neumann boundary conditions. *Journal de mathématiques pures et appliquées*, 90(2):201–227, 2008.

[2] J. Aujol, G. Gilboa, and N. Papadakis. Theoretical analysis of flows estimating eigenfunctions of one-homogeneous functionals. *SIAM Journal on Imaging Sciences*, 11(2):1416–1440, 2018.

[3] X. Bresson, T. Laurent, D. Uminsky, and J. Von Brecht. Convergence and energy landscape for Cheeger Cut clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1385–1393, 2012.

[4] X. Bresson, T. Laurent, D. Uminsky, and J. Von Brecht. Multiclass total variation clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1421–1429, 2013.

[5] X. Bresson, T. Laurent, D. Uminsky, and J. H. Von Brecht. An adaptive total variation algorithm for computing the balanced cut of a graph. *arXiv preprint arXiv:1302.2717*, 2013.

[6] M. A. Bruno, E. A. Walker, and H. H. Abujudeh. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 35(6):1668–1676, 2015.

[7] T. Bühler and M. Hein. Spectral clustering based on the graph p-laplacian. *International Conference on Machine Learning (ICML)*, 2009.

[8] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40:120–145, 2011.

[9] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning. *MIT Press*, 20(3):542–542, 2006.

[10] D. Dai and L. Van Gool. Ensemble projection for semi-supervised image classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2072–2079, 2013.

[11] L. Dodero, A. Gozzi, A. Liska, V. Murino, and D. Sona. Group-wise functional community detection through joint laplacian diagonalization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 708–715. Springer, 2014.

[12] A. Elmoataz, O. Lezoray, and S. Bougleux. Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing. *IEEE transactions on Image Processing*, 17(7):1047–1060, 2008.

[13] T. Feld, J. Aujol, G. Gilboa, and N. Papadakis. Rayleigh quotient minimization for absolutely one-homogeneous functionals. *Inverse Problems*, 2019.

[14] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *Advances in neural information processing systems (NIPS)*, pages 522–530, 2009.

[15] Y. Gao, E. Adeli-M, M. Kim, P. Giannakopoulos, S. Haller, and D. Shen. Medical image retrieval using multi-graph learning for MCI diagnostic assistance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 86–93, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[17] M. Hein, S. Setzer, L. Jost, and S. S. Rangapuram. The total variation on hypergraphs-learning on hypergraphs revisited. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2427–2435, 2013.

[18] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 7132–7141, 2018.

[19] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum. Label propagation for deep semi-supervised learning. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[20] T. Joachims. Transductive learning via spectral graph partitioning. In *International Conference on Machine Learning (ICML)*, pages 290–297, 2003.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[22] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. *International conference on Machine learning (ICML)*, 2017.

[23] Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8896–8905, 2018.

[24] S. S. Rangapuram, P. K. Mudrakarta, and M. Hein. Tight continuous relaxation of the balanced k-cut problem. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3131–3139, 2014.

[25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems (NIPS)*, pages 2234–2242, 2016.

[26] W. Shi, Y. Gong, C. Ding, Z. MaXiaoyu Tao, and N. Zheng. Transductive semi-supervised deep learning using min-max features. In *European Conference on Computer Vision (ECCV)*, pages 299–315, 2018.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

[28] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems (NIPS)*, pages 1195–1204, 2017.

[29] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[30] J. Wang, T. Jebara, and S.-F. Chang. Graph transduction via alternating minimization. In *International conference on Machine learning (ICML)*, pages 1144–1151. ACM, 2008.

[31] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2097–2106, 2017.

[32] J. Weston, F. Ratle, and R. Collobert. Deep learning via semi-supervised embedding. In *International conference on Machine learning (ICML)*, pages 1168–1175, 2008.

[33] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*, 2018.

[34] Y.-M. Zhang, K. Huang, and C.-L. Liu. Fast and robust graph-based transductive learning via minimum tree cut. In *IEEE International Conference on Data Mining*, pages 952–961, 2011.

[35] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328, 2004.

[36] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[37] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *International conference on Machine learning (ICML'03)*, pages 912–919, 2003.

Supplementary Material for: Beyond Supervised Classification: Extreme Minimal Supervision with the Graph 1-Laplacian

This supplementary material extends further details and proofs that support the content of the main paper. In particular, the proof of Proposition 2 and Proposition 3 from the main paper.

A Proofs

A.1 Proof of Proposition 2

1 For $\langle u_k, d\rangle = 0$, we have

$$\begin{aligned}\langle u_{k+1/2}, d\rangle &= \langle u_k, d\rangle + \delta t \left(R(u_k)\langle (q_k - \tilde{q}_k), d\rangle - \langle p_{k+1/2}, d\rangle\right) \\ &= \delta t R(u_k)\left(\langle q_k, d\rangle - \frac{\langle d, q_k\rangle}{\langle d, d\rangle}\langle d, d\rangle\right) \\ &= 0,\end{aligned}$$

where we used Proposition 1 in the right part of the previous relation to get $\langle p_{k+1/2}, d\rangle = 0$. We conclude with the fact that u_{k+1} is a rescaling of $u_{k+1/2}$.

2 Since H is a norm, it is absolutely one homogeneous and $q_k \in \partial H(u_k) \Rightarrow H(u_k) = \langle q_k, u_k\rangle$. Next, we observe that $J(u_k) = \sup_{p\in\partial J}\langle p, u_k\rangle \geq \langle p_{k+1/2}, u_k\rangle$ and we get

$$\begin{aligned}\langle u_{k+1/2}, u_k\rangle &= ||u_k||_2^2 + \delta t \left(R(u_k)\langle q_k - \tilde{q}_k, u_k\rangle - \langle p_{k+1/2}, u_k\rangle\right) \\ &\geq ||u_k||_2^2 + \delta t \left(J(u_k) - R(u_k)\langle \tilde{q}_k, u_k\rangle - J(u_k)\right) \\ &\geq ||u_k||_2^2 - \delta t R(u_k)\frac{\langle d, q_k\rangle}{\langle d, d\rangle}\langle d, u_k\rangle \\ &\geq ||u_k||_2^2.\end{aligned}$$

We then conclude with the fact that $\langle u_{k+1/2}, u_k\rangle \leq ||u_{k+1/2}||_2.||u_k||_2$.

3 Since $\langle u_k, d\rangle = 0$ for all k and $\tilde{q}_k = \frac{\langle d, q_k\rangle}{\langle d, d\rangle}d$, then $\langle \tilde{q}, u_{k+1/2}\rangle = \langle \tilde{q}, u_k\rangle = 0$. Next, we recall that $H(u_{k+1/2}) = \sup_{q\in\partial H.}\langle q, u_{k+1/2}\rangle \geq \langle q_k, u_{k+1/2}\rangle$. Hence we have

$$\begin{aligned} F_k(u_{k+1/2}) &\leq F(u_k) \\ \frac{1}{2\delta t}\|u_{k+1/2} - u_k\|_2^2 - R(u_k)\langle q_k, u_{k+1/2}\rangle + J(u_{k+1/2}) &\leq 0 \\ \frac{1}{2\delta t}\|u_{k+1/2} - u_k\|_2^2 + J(u_{k+1/2}) &\leq R(u_k)H(u_{k+1/2}) \\ R(u_{k+1/2}) &\leq R(u_k) \\ R(u_{k+1}) &\leq R(u_k)\end{aligned} \tag{15}$$

where the final rescaling with $\|u_{k+1/2}\|_2$ is possible since J and H are absolutely one homogeneous functions.

4 In the finite dimension setting, there exists $K_J, K_H < \infty$ such that $||p|| \leq K_J$ and $||q|| \leq K_H$ for an absolutely one homogeneous functionals J defined in (2) and a norm H.

Then one has

$$
\begin{aligned}
u_{k+1/2} &= u_k + \delta t \left(\frac{J(u_k)}{H(u_k)}(q_k - \tilde{q}_k) - p_{k+1/2} \right) \\
\|u_{k+1/2}\|_2^2 &= \langle u_k, u_{k+1} \rangle + \delta t \left(\frac{J(u_k)}{H(u_k)} \langle q_k, u_{k+1/2} \rangle - \langle p_{k+1/2} \rangle \right) \\
\|u_{k+1/2}\|_2^2 &\leq \|u_{k+1/2}\|_2 \left(\|u_k\|_2 + \delta t \left(\frac{J(u_k)}{H(u_k)} K_H + K_J \right) \right) \\
\|u_{k+1/2}\|_2 &\leq 1 + \delta t \left(\frac{J(u_0)}{H(u_0)} K_H + K_J \right).
\end{aligned}
$$

Hence from the equivalence of norms in finite dimensions, there exists $0 < \kappa < \infty)$ such that $H(u_{k+1/2} \leq \kappa$.

A.2 Proof of Proposition 3

Proof. 1 For $\langle u_k^l, d \rangle = 0$, and following point 1 of Proposition 2, we have

$$
\begin{aligned}
\langle u_{k+1/2}^l, d \rangle &= \langle u_k^l, d \rangle + \delta t \left(R(u_k^l) \langle (q_k^l - \tilde{q}_k^l), d \rangle - \langle p_{k+1/2}^l, d \rangle - \langle r_{k+1/2}^l, d \rangle \right) \\
&= -\langle r_{k+1/2}^l, d \rangle \\
&= -\langle \alpha, d \rangle,
\end{aligned}
$$

where we used the characteriation of $\mathbf{r}$ in (13). Next, as $\mathbf{u}_{k+1/2} \in C$, we have $\sum_l u_{k+1/2}^l(x) = 0$, $\forall x \in \mathcal{N}$ and obtain:

$$
\begin{aligned}
\sum_{l=1}^{L} \langle u_{k+1/2}^l, d \rangle &= -\sum_{l=1}^{L} \langle \alpha, d \rangle \\
\sum_{l=1}^{L} \sum_{x \in \mathcal{N}} u_{k+1/2}^l(x) d(x) &= -L \langle \alpha, d \rangle \\
\sum_{x \in \mathcal{N}} d(x) \left(\sum_{l=1}^{L} u_{k+1/2}^l(x) \right) &= -L \langle \alpha, d \rangle \\
0 &= \langle \alpha, d \rangle.
\end{aligned}
$$

2 We have

$$
\langle u_{k+1/2}^l, u_k^l \rangle = ||u_k^l||_2^2 + \delta t \left(R(u_k^l) \langle q_k^l - \tilde{q}_k^l, u_k^l \rangle - \langle p_{k+1/2}^l, u_k^l \rangle - \langle r_{k+1/2}^l, u_k^l \rangle \right).
$$

We follow the point 2 of Proposition 2 to first get: $\langle u_{k+1/2}^l, u_k^l \rangle \geq \|u_k^l\|_2 - \langle r_{k+1/2}^l, u_k^l \rangle$, for $i = 1 \cdots n$. Then, as $\sum_l \langle r_{k+1/2}^l, u_k^l \rangle = \langle \mathbf{r}_{k+1/2}, \mathbf{u}_k \rangle \leq \chi_C(\mathbf{u}_k) = 0$, we deduce that $\|\mathbf{u}_{k+1/2}\|_2 . \|\mathbf{u}_k\|_2 \geq \sum_l \langle u_{k+1/2}^l, u_k^l \rangle \geq \sum_l \langle u_k^l, u_k^l \rangle = ||\mathbf{u}_k||_2^2$. Next we have

$$
\|u_{k+1/2}^l\|_2^2 = \langle u_{k+1/2}^l, u_k^l \rangle + \delta t \left(R(u_k^l) \langle q_k^l - \tilde{q}_k^l, u_{k+1/2}^l \rangle - J(u_{k+1/2}^l) - \langle r_{k+1/2}^l, u_{k+1/2}^l \rangle \right).
$$

Summing on l, we get

$$
\begin{aligned}
\|\mathbf{u}_{k+1/2}\|_2^2 &\leq \|\mathbf{u}_{k+1/2}\|_2 \left(\|\mathbf{u}_k\|_2 + \delta t \left(\sum_{l=1}^{L} R(u_k^l) \|q_k^l\|_2 + \|p_{k+1/2}^l\|_2 \right) \right) \\
\|\mathbf{u}_{k+1/2}\|_2 &\leq \|\mathbf{u}_k\|_2 + \delta t \left(\sum_{l=1}^{L} \frac{J(u_k^l)}{H(u_k^l)} K_H + K_J \right) \leq 1 + \delta t K_J \left(\sum_{l=1}^{L} \frac{\|u_k^l\|_2}{H(u_k^l)} K_H + 1 \right)
\end{aligned}
$$

Notice that we defined $R(u_k^l) = 0$ for $u_k^l = 0$. As H is a norm, the equivalence of norm in finite dimensions implies that $\|u_k^l\|_2 H(u_k^l)$ is bounded by some constant $c < \infty$. We then have $\|\mathbf{u}_{k+1/2}\|_2 \leq \kappa = 1 + \delta t K_J (1 + L K_H c)$.

3 Since $\mathbf{u}_{k+1/2}$ is the global minimizer of (11), then:

$$
\begin{aligned}
F_k^L(\mathbf{u}_{k+1/2}) &\leq F_k^L(\mathbf{u}_k) \\
\frac{1}{2\delta t}\|\mathbf{u}_{k+1/2}-\mathbf{u}_k\|_2^2+\sum_{l=1}^{L} J(u_{k+1/2}^l) &\leq \sum_{l=1}^{L} R(u_k^l)\langle q_k^l-\tilde{q}_k^l, u_{k+1/2}^l\rangle \\
\frac{1}{2\delta t}\|\mathbf{u}_{k+1/2}-\mathbf{u}_k\|_2^2+\sum_{l=1}^{L} J(u_{k+1/2}^l) &\leq \sum_{l=1}^{L} R(u_k^l) H(u_{k+1/2}^l) \\
\sum_{l=1}^{L}\left(J(u_{k+1/2}^l)-\frac{J(u_k^l)}{H(u_k^l)}H(u_{k+1/2}^l)\right) &\leq -\frac{1}{2\delta t}\|\mathbf{u}_{k+1/2}-\mathbf{u}_k\|_2^2 \\
\|\mathbf{u}_{k+1/2}\|_2\sum_{l=1}^{L} H(u_{k+1}^l)\left(R(u_{k+1}^l)-R(u_k^l)\right) &\leq -\frac{1}{2\delta t}\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2 \\
\sum_{l=1}^{L} H(u_{k+1}^l)\left(R(u_{k+1}^l)-R(u_k^l)\right) &\leq -\frac{1}{2\delta t\kappa}\|\mathbf{u}_{k+1}-\mathbf{u}_k\|_2^2.
\end{aligned}
$$

□