A Unified Weight Learning Paradigm for Multi-view Learning

Lai Tian tianlai.cs@gmail.com
Feiping Nie* feipingnie@gmail.com
Xuelong Li li@nwpu.edu.cn
School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.

Abstract

Learning a set of weights to combine views linearly forms a series of popular schemes in multi-view learning. Three weight learning paradigms, *i.e.*, Norm Regularization (NR), Exponential Decay (ED), and p-th Root Loss (pRL), are widely used in the literature, while the relations between them and the limiting behaviors of them are not well understood yet. In this paper, we present a Unified Paradigm (UP) that contains the aforementioned three popular paradigms as special cases. Specifically, we extend the domain of hyper-parameters of NR from positive to real numbers and show this extension bridges NR, ED, and pRL. Besides, we provide detailed discussion on the weights sparsity, hyper-parameter setting, and counterintuitive limiting behavior of these paradigms. Furthermore, we show the generality of our technique with examples in Multi-Task Learning and Fuzzy Clustering. Our results may provide insights to understand existing algorithms better and inspire research on new weight learning schemes. Numerical results support our theoretical analysis.

1 Introduction

In recent years, several methods to analyze multi-view data have been proposed. These methods learn from data by considering the diversity and complementary of different views. We can easily obtain data with multiple views from multiple sources or from different feature subsets [XTX13]. For example, we can identify a person by face, fingerprint, signature or iris with information obtained from multiple sources and we can represent an image by its color or texture features, which can be seen as different feature subsets of the image.

A direct way to integrate multi-view data is to concatenate all the feature vectors into a long one and perform single view algorithm on the long vectors. But this concatenation causes overfitting in the case of a small size of training sample and is not physically meaningful since each view has a specific statistical property [XTX13]. Smarter strategies should be considered to fully exploit the multiple views of multi-view data. Several works have been done and many of them prefer to construct a similarity graph for every view and then linearly combine these graphs to build a unified one [KM13, CNCH13, NCL17]. For instance, we use linear combination $\mathbf{S} = \sum_{i=1}^{n} \alpha_i \mathbf{A}^{(i)}$ to learn the new similarity graph $\mathbf{S} \in \mathbb{R}^{n \times n}$ from $\{\mathbf{A}^{(i)}\}$, where $\mathbf{A}^{(i)} \in \mathbb{R}^{n \times n}$ is the affinity graph built from the i-th view [NLL16]. Recently, some work on multi-view learning under the PAC-Bayes framework has been done [GMA18, Goy18, GMGA17, SSTM17], in which the normalized combination weights are supposed to be the posterior distribution.

Since the weights $\{\alpha_i\}_{i=1}^{n}$ in above settings dominate the performance of algorithms, better and smarter strategy to set these weights is of cardinal significance. Due to predetermined and fixed weights are hard to tune, not adaptive to data, and often have unsatisfied performance, we often prefer to learn these weights out directly from data. There are several paradigms in the literature to learn these weights and detailed discussions are provided in the next subsection.

1.1 Problem Statement

The basic weight learning framework for linear combination of views considered in this paper is that

$$\min_{x \in \mathcal{S}, \alpha} \sum_{i=1}^{n} \alpha_i f_i(x) \quad s.t. \quad \sum_{i=1}^{n} \alpha_i = 1, \alpha_i \geq 0$$

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

where $f_i(x)$ is a problem specific function and $\mathcal{S}$ is the domain of x. In multi-view learning, $f_i(x)$ should be the loss function for the i-th view, *e.g.*, the quadratic smoothness penalty $Tr(\mathbf{F}^\top \mathbf{L}^{(i)} \mathbf{F})$ and x should be the task dependent variable, *e.g.*, spectral embedding $\mathbf{F} \in \mathbb{R}^{n\times k}$, where $\mathbf{L}^{(i)}$ is the Graph Laplacian [Chu97] of the i-th view and k is the number of clusters. Often, a probability simplex constraint $\sum_{i=1}^n \alpha_i = 1, \alpha_i \geq 0$ is applied to the weights $\{\alpha_i\}_{i=1}^n$ to make them normalized and have somewhat probability meaning. Sometimes one may concern the objective function with additional regularization term for x, *e.g.*, $\sum_{i=1}^n \alpha_i f_i(x) + \lambda \mathcal{R}(x)$. In that case, use the new domain set $\mathcal{S}' := \{x | x \in \mathcal{S}, \mathcal{R}(x) \leq \mu\}$ and the regularization term vanishes.

However, above basic framework has a trivial solution with respect to $\{\alpha_i\}_{i=1}^n$ [KM13], that is $\alpha_i = 1$ if $i = \arg\min_k f_k(x)$ and it is 0 otherwise (to see it, note that the objective function is linear with respect to $\{\alpha_i\}_{i=1}^n$ and use the Proposition B.19 of [Ber99]). This trivial solution makes the basic framework inapplicable in practice, since the learned weights is too sparse to make full use of the multi-view information contained in $\{f_i(x)\}_{i=1}^n$.

To overcome this *over-sparse* problem, several approaches were proposed:

Exponential Decay. It has been widely used [HYZ+18, LJL+15, LJW+17, WCLC15, XWL16, ZHWZ16, XLW+15, TL12, TL10, XTMZ10] to obtain dense solution by an additional exponential decay factor γ.

$$\min_{x\in\mathcal{S},\alpha} \sum_{i=1}^n \alpha_i^\gamma f_i(x) \quad \text{s.t.} \quad \sum_{i=1}^n \alpha_i = 1, \alpha_i \geq 0,$$

where $\gamma \geq 1$ is the decay hyper-parameter to tune. Note that when $\gamma = 1$, the exponential decay reduces to the over-sparse problem.

Example 1. *The following problem is a multi-view embedding model, termed Multiview Spectral Embedding (MSE) [XTMZ10], which is not recently proposed but simple enough to demonstrate the Exponential Decay paradigm:*

$$\min_{\substack{\mathbf{F}^\top\mathbf{F}=\mathbf{I}\\ \alpha\geqslant 0, \alpha^\top \mathbf{1}=1}} \sum_{i=1}^c \alpha_i^\gamma Tr\left(\mathbf{F}^\top \mathbf{L}^{(i)} \mathbf{F}\right),$$

where c is the number of views, $\mathbf{L}^{(i)} \in \mathbb{R}^{n\times n}$ is the Laplacian of the i-th view and $\mathbf{F} \in \mathbb{R}^{n\times k}$ (k is the number of clusters) is the learned spectral embedding. This embedding $\mathbf{F}$ is a new representation of original data and can be used for clustering or retrieval.

Norm Regularization. Another popular sparse restraining technique is to regularize the linear combination with a ℓ_2 norm. This paradigm is also quite popular among researchers [KM13, ZNHY17, ZHJ+17, CRNA14, XWL16, KR15]. While the ℓ_2 norm is widely used, it can be generalized to ℓ_q norm with $q \geq 1$:

$$\min_{x\in\mathcal{S},\alpha} \sum_{i=1}^n \alpha_i f_i(x) + \lambda \alpha_i^q \quad \text{s.t.} \quad \sum_{i=1}^n \alpha_i = 1, \alpha_i \geq 0,$$

where $\lambda \geq 0$ controls the sparsity of the solution $\{\alpha_i\}_{i=1}^n$. Note that when $q = 1$, the norm regularization reduces to the over-sparse problem.

Example 2. *A semi-supervised Learning method for multi-view data, named Sparse Multiple Graph Integration (SMGI) [KM13], fits into the Norm Regularization paradigm. SMGI learns weighs to linearly combine multiple graphs from given labels. The objective of SMGI is*

$$\min_{\substack{\mathbf{F},\alpha\geqslant 0\\ \alpha^\top\mathbf{1}=1}} \sum_{i=1}^c \left(\frac{\alpha_i}{Z_i} Tr\left(\mathbf{F}^\top \mathbf{L}^{(i)} \mathbf{F}\right) + \lambda_1 \alpha_i^2\right) + \lambda_2 \|\mathbf{F} - \mathbf{Y}\|_F^2,$$

where $\mathbf{Y} \in \mathbb{R}^{n\times k}$ is the given labels. The "sparse" in the name of SMGI refers to the sparsity of optimal weights $\{\alpha_i\}_{i=1}^c$, which will be discussed in Section 4.1.

***p*-th Root Loss.** A relatively new weight learning paradigm [NLL16, NCL17, SWF+17, XNL+17, NTL18] considers the p-th root of $\{f_i(x)\}_{i=1}^n$, where the weights $\{\alpha_i\}_{i=1}^n$ are implicitly and adaptively defined. Formally, the p-th Root Loss paradigm solves $\min_x \sum_{i=1}^n \sqrt{f_i(x)}$, or more generally, for $0 < p \leq 1$,

$$\min_x \sum_{i=1}^n f_i(x)^p.$$

It is shown in [NLL16] that above optimization problem can be viewed as an adaptively weighted problem, $\min_{x\in\mathcal{S},\alpha} \sum_{i=1}^n \alpha_i f_i(x)$ with $\alpha_i = \frac{p}{f_i(x)^{1-p}}$.

Example 3. *The framework Auto-weighted Multiple Graph Learning (AMGL) proposed in [NLL16] is aim for both multi-view clustering and multi-view semi-supervised learning. For simplicity, we only present the optimization problem for clustering, that is*

$$\min_{\mathbf{F}^\top\mathbf{F}=\mathbf{I}} \sum_{i=1}^c \sqrt{Tr\left(\mathbf{F}^\top \mathbf{L}^{(i)} \mathbf{F}\right)}.$$

In the analysis of [NLL16], the authors introduce dynamical and implicit weights $\{\alpha_i\}_{i=1}^n$, which are defined by $\alpha_i := 1/\left(2\sqrt{\mathbf{F}^\top \mathbf{L}^{(i)} \mathbf{F}}\right)$. With these implicit weights, the authors show the objective function of AMGL can be reformulated as $\sum_{i=1}^n \alpha_i Tr\left(\mathbf{F}^\top \mathbf{L}^{(i)} \mathbf{F}\right)$. But it is somewhat unnatural and vague to analyze with such implicit weights. It is of interest to seek an explicit linear combination expression for it.

While these three weight learning paradigms are widely used in the literature and have shown their effectiveness on multi-view learning, it is of interest to ask following questions:

1. Are there any connections between these three paradigms?
2. Do we really need three, rather than one, distinct paradigms to learn these weights?

1.2 Contributions and Paper Outline

In this paper, we address these questions by establishing the connections between aforementioned three weight learning schemes and presenting a unified weights learning paradigm for multi-view learning. In the new paradigm, the above three schemes, ED, NR, and pRL, are framed as special cases. The main contributions and paper outline are listed as follows:

- We present a Unified Paradigm (UP) contains NR, ED, and pRL as special cases and show connections between them in Section 2.
- We provide some interesting observations concerning the relation of weights sparsity and the hyperparameter, the counterintuitive limiting behavior of ED, and some interesting reformulations of Fuzzy C-Means in Section 4.
- We show the generality of our technique by providing Multi-Task Learning example that can be fitted into the proposed paradigm in Section 4.4.
- Numerical results are given to validate our theoretical results in Section 5.

Notations. Most of the notations used in this paper are standard. We use $\|\mathbf{x}\|_p = (\sum_{i=1}^n x_i^p)^{1/p}$ for ℓ_p norm and omit the subscript p when $p = 2$. Set of indexed elements is denoted by $\{x_i\}_{i=1}^n$. The trace of a matrix is denoted by $Tr(\mathbf{X}) = \sum_{i=1}^n x_{ii}$. In the whole paper, we will assume $f_i(x) \geq 0, \forall x$, which is often the case in the multi-view learning literature.

2 The Unified Paradigm

In this section, we first introduce some useful notions and then present the newly proposed paradigm.

2.1 Preliminaries

We first introduce some notions.

Definition 1 (x-partial equivalence)**.** *We write*

$$\min_{x,\alpha} f_1(x,\alpha) \simeq_x \min_{x,\alpha} f_2(x,\alpha)$$

if and only if

$$\arg_x \min_{x,\alpha} f_1(x,\alpha) = \arg_x \min_{x,\alpha} f_2(x,\alpha).$$

Remark 1. *It is easy to see that the partial equivalence is transitive with respect to x, which means that we can say $\min_{x,\alpha} f_1(x,\alpha) \simeq_x \min_{x,\alpha} f_3(x,\alpha)$, if we have $\min_{x,\alpha} f_1(x,\alpha) \simeq_x \min_{x,\alpha} f_2(x,\alpha)$ and $\min_{x,\alpha} f_2(x,\alpha) \simeq_x \min_{x,\alpha} f_3(x,\alpha)$.*

Remark 2. *The motivation for us to use the partial equivalence is that sometimes we actually do not care about the exact value of learned weights, since the learned x rather than $\{\alpha_i\}_{i=1}^n$ is the final learning result. Suppose two multi-view learning algorithms get the same linear combined Laplacian $\mathbf{L} = \sum_{i=1}^n \alpha_i \mathbf{L}^{(i)}$ with different $\{\alpha_i\}_{i=1}^n$. It is difficult to tell which one learns better. Thus, partial equivalence is quite suitable for such situations.*

Definition 2 (Power mean)**.** *Denote the order p generalized mean of set $\{x_i\}_{i=1}^n$ by*

$$\mathbb{M}_p(\{x_i\}) := \sqrt[p]{\frac{1}{n}\sum_{i=1}^n x_i^p}.$$

The family of power means (a.k.a. generalized means) include the classical Pythagorean means, *i.e.*, Arithmetic ($p = 1$), Geometric ($p = 0$), and Harmonic means ($p = -1$), as special cases. The *Power mean inequality*, which also contains the Pythagorean means inequality, is as follows.

Lemma 1 (Power mean inequality [Bul13])**.** *Given $p < q$, we have*

$$\sqrt[p]{\frac{1}{n}\sum_{i=1}^n x_i^p} = \mathbb{M}_p(\{x_i\}) \leq \mathbb{M}_q(\{x_i\}) = \sqrt[q]{\frac{1}{n}\sum_{i=1}^n x_i^q}.$$

Specially,

$$\mathbb{M}_{-\infty}(\{x_i\}) = \min_i\{x_i\}, \quad \mathbb{M}_{+\infty}(\{x_i\}) = \max_i\{x_i\},$$

$$\mathbb{M}_0(\{x_i\}) = \prod_{i=1}^n \sqrt[n]{x_i}.$$

2.2 The Unified Paradigm

Given $\lambda, q \in \mathbb{R}$, the new unified paradigms for multi-view learning is

$$\min_{\substack{x\in\mathcal{S},\alpha\succcurlyeq 0\\ \alpha^\top \mathbf{1}=1}} \sum_{i=1}^n \alpha_i f_i(x) + \lambda \alpha_i^q.$$

It seems that our new unified paradigm has no difference from the NR. But note that in the new model,

the hyper-parameters q and λ take values from the whole set of real numbers, which means that one can set $\lambda < 0$ or $q < 0$. This may seem weird at the first glance, since it is unusual to set the hyper-parameter λ to be negative. It will be clear in the next subsection that this domain extension of hyper-parameters bridges NR, ED, and pRL.

Though λ and q can be chosen as arbitrary real numbers, we will not consider the situations that make the optimization problem non-convex with respect to $\{\alpha_i\}_{i=1}^n$, *e.g.*, $\lambda > 0$ and $0 < q < 1$. The reader may have concerns here since regularization with $\|\cdot\|_q, 0 < q < 1$ is widely used in the literature. But note that such kind of regularization is often named *sparsity-inducing* penalty [BJM+12], whose main purpose is to promote the sparsity of the weights $\{\alpha_i\}_{i=1}^n$. On the contrary, *over-sparse* is the main obstacle to be overcome in our problem setting. Therefore, we use the opposite, *i.e.*, $\lambda < 0$ and $0 < q < 1$, which leads to a convex optimization problem with respect to $\{\alpha_i\}_{i=1}^n$ and has the *density-inducing* effect.

2.3 Main Theory

In this subsection, we establish connections between NR, ED, pRL, and the newly proposed unified paradigm, that is, different hyper-parameter regions make UP become NR, ED, or pRL. Our main results can be framed into Table 1 and summarized into following theorem:

Theorem 1. *It holds that*

1. *If $\gamma > 1, q = \frac{1}{\gamma}$, then there exists $\lambda < 0$ such that*
$$\min_{\substack{x\in\mathcal{S},\alpha\succcurlyeq 0\\ \alpha^\top \mathbf{1}=1}} \sum_{i=1}^n \alpha_i f_i(x) + \lambda \alpha_i^q \simeq_x \min_{\substack{x\in\mathcal{S},\alpha\succcurlyeq 0\\ \alpha^\top \mathbf{1}=1}} \sum_{i=1}^n \alpha_i^\gamma f_i(x).$$
2. *If $0 < p < 1, \frac{1}{p} + \frac{1}{q} = 1$, then there exists $\lambda > 0$ such that $q < 0$ and*
$$\min_{\substack{x\in\mathcal{S},\alpha\succcurlyeq 0\\ \alpha^\top \mathbf{1}=1}} \sum_{i=1}^n \alpha_i f_i(x) + \lambda \alpha_i^q \simeq_x \min_{x\in\mathcal{S}} \sum_{i=1}^n f_i(x)^p.$$
3. *Given $q \neq 0$, there exists λ with $q < 0 < \lambda$ or $\lambda < 0 < q < 1$ such that*
$$\min_{\substack{x\in\mathcal{S},\alpha\succcurlyeq 0\\ \alpha^\top \mathbf{1}=1}} \sum_{i=1}^n \alpha_i f_i(x) + \lambda \alpha_i^q \simeq_x \min_{x\in\mathcal{S}} \mathbb{M}_{\frac{q}{q-1}}(\{f_i(x)\}).$$

Remark 3. *Several remarks can be made.*

- *The point 1 in Theorem 1 establishes connections between ED and UP, which indicates that an ED-type model can be rewritten as its equivalent regularization form. To the best of our knowledge, this regularization-type equivalence is previously unknown and lots of existing ED model [HYZ+18, LJL+15, LJW+17, WCLC15, XWL16, ZHWZ16, XLW+15, TL12, TL10] can be framed into the regularization framework accordingly. Meanwhile, this connection reveals that some not well-known regularization term, e.g, $-\sqrt{\alpha_i}$ has been widely used implicitly, e.g., in the form of ED with $\gamma = 2$.*
- *The connection between pRL and UP is revealed in the point 2 of Theorem 1, which indicates that the composition of the p-th root and original loss, i.e., $\sum_{i=1}^n f_i(x)^p$, has an equivalent regularization form, that is, the linear combination of original loss and an additional special regularization term, e.g., $\sum_{i=1}^n \sqrt{f_i(x)} \simeq_x \sum_{i=1}^n \alpha_i f_i(x) + \lambda \alpha_i^{-1}$. Meanwhile, this equivalent regularization form provides an explicit weight learning model for the dynamic weight interpretation in [NLL16] (see Example 3).*
- *In the point 3 of Theorem 1, it shows that when $q < 0 < \lambda$ or $\lambda < 0 < q < 1$, UP is equivalent to a power mean minimization problem with specific order. This reformulation allows one to use power mean inequality (Lemma 1) to investigate the limit behavior of UP, ED, and pRL, while in their original form the limiting behavior may not be easy to see or even misleading. For example, a commonly believe in the literature [TL12, ZDF17] is that when $\gamma \to +\infty$, ED will become Arithmetic Mean minimization. But in Section 4.2 we will show it should be Geometric Mean rather than Arithmetic Mean.*

3 Proof Sketch

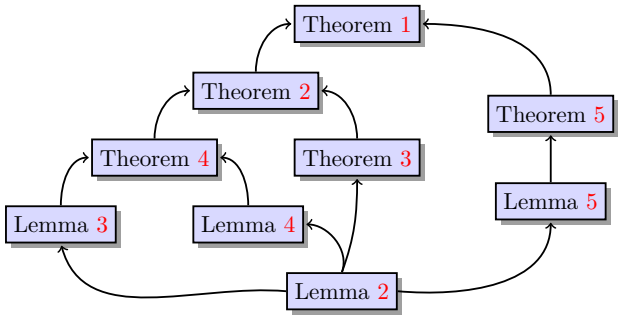

Figure 1: Flowchart of the proofs.

In this section, we present the proof sketch for the Theorem 1. The main ideas and technical lemmas are given while the rigorous technical proofs of them are deferred to the appendix.

Table 1: Summarization of relations between ED, NR, pRL, and the Unified Paradigm (UP).

$\sum_i \alpha_i f_i(x) + \lambda\alpha_i^q$	$\lambda < 0$	$\lambda > 0$
$q > 1$	Non-Convex	Norm Regularization ($\sum_{i=1}^n \alpha_i f_i(x) + \lambda\alpha_i^q$)
$0 < q < 1$	Exponential Decay ($\sum_{i=1}^n \alpha_i^\gamma f_i(x)$)	Non-Convex
$q < 0$	Non-Convex	p-th Root Loss ($\sum_{i=1}^n f_i(x)^p$)

The main difficulty to analyze the optimization problem

$$\min_{x,\alpha^\top \mathbf{1}=1,\alpha\succcurlyeq 0} \sum_{i=1}^n \alpha_i f_i(x) + \lambda\alpha_i^q$$

is the probability constraint $\alpha^\top \mathbf{1} = 1$. Our strategy is that we first show the results hold without $\alpha^\top \mathbf{1} = 1$, then show that there exists λ such that putting the constraint back does not change the results.

Instead of analyzing the unified model directly, we first consider a less constrained version by removing the constraint $\sum_{i=1}^n \alpha_i = 1$.

Theorem 2 (Less constrained)**.** *It holds that*

- *If $\gamma > 1, q = \frac{1}{\gamma}$, and $\lambda < 0$, then $0 < q < 1$ and*

$$\min_{x\in\mathcal{S},\alpha\succcurlyeq 0} \sum_{i=1}^n \alpha_i f_i(x) + \lambda\alpha_i^q \simeq_x \min_{\substack{x\in\mathcal{S},\alpha\succcurlyeq 0\\ \alpha^\top\mathbf{1}=1}} \sum_{i=1}^n \alpha_i^\gamma f_i(x).$$

- *If $0 < p < 1, \frac{1}{p} + \frac{1}{q} = 1$, and $\lambda > 0$, then $q < 0$ and*

$$\min_{x\in\mathcal{S},\alpha\succcurlyeq 0} \sum_{i=1}^n \alpha_i f_i(x) + \lambda\alpha_i^q \simeq_x \min_{x\in\mathcal{S}} \sum_{i=1}^n f_i(x)^p.$$

Theorem 2 contains two equivalence relations. One of these relations is easy to establish while the other needs more steps. But for both of these relations, a reformulation of the unified model will be helpful.

Lemma 2. *Let*

$$C = \left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} + \lambda\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{q}{q-1}}.$$

Then, we have

$$\min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^n \alpha_i f_i(x) + \lambda\alpha_i^q \simeq_x \min_x C \cdot \sum_{i=1}^n f_i(x)^{\frac{q}{q-1}}.$$

Evaluate C in Lemma 2 with $0 < p < 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. One of the equivalence relation in Theorem 2 just follows:

Theorem 3. *Given $0 < p < 1, \frac{1}{p} + \frac{1}{q} = 1$, then $q < 0 < \lambda$ and*

$$\min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^n \alpha_i f_i(x) + \lambda\alpha_i^q \simeq_x \min_x \sum_{i=1}^n f_i(x)^p.$$

To get the other part of Theorem 2, we first establish the connection between the unified model and power mean.

Lemma 3. *Given $\lambda < 0 < q < 1$, then $\sum_{i=1}^n \alpha_i f_i(x) + \lambda\alpha_i^q$ is convex with respect to $\{\alpha_i\}_{i=1}^n$ and*

$$\min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^n \alpha_i f_i(x) + \lambda\alpha_i^q \simeq_x \min_x \mathbb{M}_c(\{f_i(x)\}),$$

where $c = \frac{q}{q-1}$.

Then, we show that ED is also partial equivalent to the power mean minimization.

Lemma 4. *Let $\gamma > 1$. Then,*

$$\min_{\alpha^\top\mathbf{1}=1,\alpha\succcurlyeq 0} \sum_{i=1}^n \alpha_i^\gamma f_i(x) \simeq_x \min_x \mathbb{M}_{\frac{1}{1-\gamma}}(\{f_i(x)\}).$$

Combining Lemma 3 and 4 with $q = \frac{1}{\gamma}$, we have

Theorem 4. *Let $\gamma > 1, q = \frac{1}{\gamma}, \lambda < 0$. Then,*

$$\min_{x,\alpha} \sum_{i=1}^n \alpha_i^\gamma f_i(x) \simeq_x \min_{x,\alpha} \sum_{i=1}^n \alpha_i f_i(x) + \lambda\alpha_i^q.$$

which justifies the second part of Theorem 2.

To get Theorem 1, we need to consider the effectiveness of the constraint $\sum_{i=1}^n \alpha_i = 1$. We first make a key observation on the ineffectiveness of $|\lambda|$ in the following lemma, which may be interesting on its own:

Lemma 5 (Ineffectiveness of $|\lambda|$)**.** *Given $q < 0 < \lambda$ or $\lambda < 0 < q < 1$, we have*

$$\min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^n \alpha_i f_i(x) + \lambda\alpha_i^q \simeq_x \min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^n \alpha_i f_i(x) + sgn(\lambda)\cdot\alpha_i^q.$$

It seems unusual that the absolute value of hyper-parameter λ has no effect on the optimal x. But note that the objective function of ED or pRL has only one hyper-parameter to tune, *i.e.*, γ in ED or p in pRL, while the corresponding form of the less constrained unified model has two, *i.e.*, λ and q. Therefore, if the equivalence between ED, pRL, and the less constrained unified model really holds, the ineffectiveness of $|\lambda|$ is reasonable enough.

Finally, we complete the chain of proofs for Theorem 1 by showing that there exists λ such that the probabilistic simplex constraint $\sum_{i=1}^n \alpha_i = 1$ can be safely put back.

Theorem 5 (Ineffectiveness of $\alpha^\top \mathbf{1} = 1$)**.** *Given $q < 0 < \lambda$ or $\lambda < 0 < q < 1$, there exists λ' such that*

$$\min_{x,\alpha \succcurlyeq 0} \sum_{i=1}^n \alpha_i f_i(x) + \lambda \alpha_i^q \simeq_x \min_{x,\alpha \succcurlyeq 0, \alpha^\top \mathbf{1}=1} \sum_{i=1}^n \alpha_i f_i(x) + \lambda' \alpha_i^q.$$

Putting all together proves Theorem 1.

4 Discussion

In this section, we provide interesting observations on the relation of weights sparsity and the hyper-parameter, the counterintuitive limiting behavior of ED, and some interesting reformulations of Fuzzy C-Means.

4.1 Sparsity in NR

Among the three components of UP, one cannot use ED or pRL to obtain sparse $\{\alpha_i\}_{i=1}^n$. That is to say, for ED and pRL, we always have $\forall i = 1, \ldots, n$, $\alpha_i > 0$. But for NR, if we set hyper-parameters λ and q carefully, we will have sparse $\{\alpha_i\}_{i=1}^n$ [KM13].

Thus, it is of great interest to set proper hyper-parameters to obtain desired sparsity. For example, if we want $\|\alpha\|_0 = k$, what is the proper γ and q that makes the optimal weights k-sparse? In this subsection, we provide a recommendation strategy on the setting of λ and q.

For simplicity, consider following problem with fixed constants $\{x_i\}_{i=1}^n$ and assume $x_1 \le x_2 \le \cdots \le x_n$:

$$\min_{\alpha \succcurlyeq 0, \alpha^\top \mathbf{1}=1} \sum_{i=1}^n \alpha_i x_i + \lambda \alpha_i^q,$$

where $\lambda > 0$ and $q > 1$. We have following relation between sparsity and the hyper-parameters.

Theorem 6. *If hyper-parameters λ and q satisfy*

$$\frac{1}{q}\left(\sum_{i=1}^k (x_k - x_i)^{\frac{1}{q-1}}\right)^{q-1} \le \lambda \le \frac{1}{q}\left(\sum_{i=1}^k (x_{k+1} - x_i)^{\frac{1}{q-1}}\right)^{q-1},$$

then the optimal α^* of above problem has $\|\alpha^*\|_0 = k$.

Remark 4. *Theorem 6 is a generalization of the result in [NWH14] which is only valid for $q = 2$. If we set $q = 2$, our result is consistent with [NWH14]. It is notable that when $q = 2$, the optimal Lagrange multiplier for the equation constraint has a closed-form expression, which was used in the proof of [NWH14]. But for the general case $q > 1$ in our case, there is no closed-form optimal multiplier expression.*

Remark 5. *It is notable that in practice, we cannot know $\{x_i\}_{i=1}^n$ in advance, since it usually has $x_i = f_i(x^*)$. Therefore, one can use a heuristics to update the hyper-parameter γ iteratively. Specifically, update $\gamma^{(t)}$ with the $x^{(t)}$ from the last round.*

4.2 ED and pRL Revisited

While the Exponential Decay technique has been widely used in the literature, the limiting behavior of it has not been well understood. Intuitively, one might conjecture [TL12, ZDF17] that the learned $\{\alpha_i^\gamma\}_{i=1}^n$ tends to be asymptoticly equal when $\gamma \to +\infty$ and ED solves the Arithmetic Mean, *i.e.*, $\min_x \frac{1}{n}\sum_{i=1}^n f_i(x)$ in that case. In this subsection, we show that this conjecture is false. Counterintuitively, when $\gamma \to +\infty$, solving ED is equivalent to solve the Geometric Mean of $\{f_i(x)\}_{i=1}^n$ rather than Arithmetic Mean. Formally, we have:

Corollary 1. *Given $\gamma \to +\infty$, we have*

$$\min_{x,\alpha} \sum_{i=1}^n \alpha_i^\gamma f_i(x) \simeq_x \min_x \prod_{i=1}^n \sqrt[n]{f_i(x)}.$$

Proof. From Lemma 4, we know

$$\min_{x,\alpha} \sum_{i=1}^n \alpha_i^\gamma f_i(x) \simeq_x \min_x \mathbb{M}_{\frac{1}{1-\gamma}}(\{f_i(x)\}).$$

When $\gamma \to +\infty$, we have $\frac{1}{1-\gamma} \to 0^-$, which indicates that

$$\min_{x,\alpha} \sum_{i=1}^n \alpha_i^\gamma f_i(x) \simeq_x \min_x \mathbb{M}_0(\{f_i(x)\}).$$

Using the power mean inequality in Lemma 1, we get $\mathbb{M}_0(\{f_i(x)\}) = \prod_{i=1}^n \sqrt[n]{f_i(x)}$, which completes the proof. □

It is easy to see from Corollary 1 that when $\gamma \to \infty$, generally, the learned $\{\alpha_i^\gamma\}_{i=1}^n$ are not asymptoticly equal. Similarly, one can get the limiting behavior of pRL when $p \to 0$ as follows:

Corollary 2. *Given $p \to 0$, we have*

$$\min_{x,\alpha} \sum_{i=1}^n f_i(x)^p \simeq_x \min_x \prod_{i=1}^n \sqrt[n]{f_i(x)}.$$

Proof. When $0 < p < 1$, it is easy to see

$$\min_{x,\alpha} \sum_{i=1}^n f_i(x)^p \simeq_x \min_{x,\alpha} \left(\frac{1}{n}\sum_{i=1}^n f_i(x)^p\right)^{\frac{1}{p}} \simeq_x \min_x \mathbb{M}_p(\{f_i(x)\}).$$

When $p \to 0$, we have $\min_{x,\alpha} \sum_{i=1}^n f_i(x)^p \simeq_x \min_x \mathbb{M}_0(\{f_i(x)\})$. Using the power mean inequality in Lemma 1, we get $\mathbb{M}_0(\{f_i(x)\}) = \prod_{i=1}^n \sqrt[n]{f_i(x)}$, which completes the proof. □

From above two corollaries, we know when $\gamma \to +\infty, p \to 0$, ED and pRL solve the Geometric Mean of $\{f_i(x)\}_{i=1}^n$. But sometimes minimizing the Geometric Mean is inconvenient. Following corollary shows the equivalence of the additional logarithmic loss and limiting ED and pRL.

Corollary 3. *Given $\gamma \to +\infty, p \to 0$, we have*

$$\min_{x,\alpha} \sum_{i=1}^n \alpha_i^\gamma f_i(x) \simeq_x \min_x \sum_{i=1}^n \log f_i(x) \simeq_x \min_{x,\alpha} \sum_{i=1}^n f_i(x)^p.$$

Proof. When $\gamma \to +\infty, p \to 0$, from Corollary 2 and 1, we know

$$\min_{x,\alpha} \sum_{i=1}^n \alpha_i^\gamma f_i(x) \simeq_x \min_x \mathbb{M}_0(\{f_i(x)\}) \simeq_x \min_{x,\alpha} \sum_{i=1}^n f_i(x)^p.$$

Notice that

$$\min_x \mathbb{M}_0(\{f_i(x)\}) \simeq_x \min_x \textstyle\sum_{i=1}^n \log f_i(x),$$

which completes the proof. □

Remark 6. *Note that in the Corollary 3 the logarithm of $\sum_{i=1}^n \log f_i(x)$ is written without base. Actually, using the logarithmic identity $\log_a b = \frac{\log_c b}{\log_c a}$ to change the base, one can easily validate that any base for the logarithm makes Corollary 3 hold.*

4.3 K-Means and Fuzzy C-Means

K-Means [FHT01] and Fuzzy C-Means [BEF84] are popular models for clustering analysis. They both aim to partition n observations into k clusters. The difference between them is that Fuzzy C-Means performs *soft clustering*, in which each data point can belong to multiple clusters, while K-Means performs *hard clustering*. The optimization objective of K-Means and Fuzzy C-Means can be written as [BEF84]:

$$\text{(K-Means):} \quad \min_{\substack{\mathbf{m},\alpha \succcurlyeq 0 \\ \alpha^\top \mathbf{1}=1}} \sum_{i=1}^n \sum_{j=1}^k \alpha_{ij} \|x_i - m_j\|^2,$$

$$\text{(Fuzzy C-Means):} \quad \min_{\substack{\mathbf{m},\alpha \succcurlyeq 0 \\ \alpha^\top \mathbf{1}=1}} \sum_{i=1}^n \sum_{j=1}^k \alpha_{ij}^2 \|x_i - m_j\|^2.$$

Using the relation between ED and the power mean in Lemma 4, the objective of K-Means ($\gamma = 1$ in ED) and Fuzzy C-Means ($\gamma = 2$ in ED) can be reformulated as

$$\begin{aligned}
\text{(K-Means):} \quad & \min_{\mathbf{m}} \sum_{i=1}^n \mathbb{M}_{-\infty}\left(\left\{\|x_i - m_j\|^2\right\}_{j=1}^k\right) \\
&\simeq_x \min_{\mathbf{m}} \sum_{i=1}^n \min\left\{\|x_i - m_j\|^2\right\}_{j=1}^k, \\
\text{(Fuzzy C-Means):} \quad & \min_{\mathbf{m}} \sum_{i=1}^n \mathbb{M}_{-1}\left(\left\{\|x_i - m_j\|^2\right\}_{j=1}^k\right) \\
&\simeq_x \min_{\mathbf{m}} \sum_{i=1}^n \left(\frac{1}{\sum_{j=1}^k 1/\|x_i - m_j\|^2}\right).
\end{aligned}$$

It is easy to see that K-Means minimizes the minimal squared distance between every data point x_i and all k prototypes, while Fuzzy C-Means minimizes the harmonic mean of these distances. That is to say, Fuzzy C-Means is equivalent to K-Harmonic Means [ZHD99].

Remark 7. *It is notable that the equivalence between Fuzzy C-Means and K-Harmonic Means has been discovered in [ZF10]. But our results is more general and can provide more equivalent forms. For example, simply using the point 1 in Theorem 2, one can get a regularization form Fuzzy C-Means, which is previously unknown. Specifically, following problem is equivalent to Fuzzy C-Means:*

$$\textit{(RegFCM):} \quad \min_{\mathbf{m},\alpha \succcurlyeq 0} \sum_{i=1}^n \sum_{j=1}^k \alpha_{ij} \|x_i - m_j\|^2 - \sqrt{\alpha_{ij}}$$

where $-\sqrt{\alpha_{ij}}$ is the regularization term and we use Lemma 5 to eliminate the hyper-parameter λ. Generally, for objective function with form $\sum_{i=1}^n \alpha_i^\gamma f_i(x)$, its regularization form is $\sum_{i=1}^n \alpha_i f_i(x) - \sqrt[q]{\alpha_i}$ with $q = \frac{1}{\gamma}$.

4.4 Further Applications

As you may see in Section 4.3, our results are not limited in multi-view learning and they can be used in many topics whose objectives are linear combination of fixed functions over the space of inputs and weights. In this subsection, we provide example in multi-task learning [NHL18] that can be framed with the Unified Paradigm.

Example 4 (Multi-Task Feature Learning)**.** *Inspired by the Calibrated Multivariate Regression (CMR) [LWZ14], a multi-task feature learning method with calibration was proposed in [GZFY14], which calibrates each task by considering the different noise levels of all tasks. The objective is given by*

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times c}} \sum_{i=1}^c \|\mathbf{X}_i \mathbf{w}_i - \mathbf{y}_i\| + \left(\lambda_1 \|\mathbf{W}\|_{1,2} + \frac{\lambda_2}{2} \|\mathbf{W}\|_F^2\right),$$

where $\mathbf{X}_i \in \mathbb{R}^{n_i \times d}$ is the data matrix and $\mathbf{y}_i \in \mathbb{R}^{n_i}$ is the response vector for the i-th task, $\mathbf{W} \in \mathbb{R}^{d \times c}$ is

Table 2: Experimental results of UP on MSRCv1.

	Unified Paradigm									
λ	0.6000	1.0000	3.0000	9.0000	26.6313	-27.4316	-10.0000	-4.0000	-3.0000	-30.0000
q	-0.8100	-0.6100	-0.4100	-0.2100	-0.0100	0.0100	0.2100	0.4100	0.6100	0.8100
NMI	0.7561	0.7521	0.7561	0.7561	0.7544	0.7544	0.7612	0.7539	0.7544	0.7612

Table 3: Experimental results of AMGL and MSE on MSRCv1.

	AMGL ($q = p/(p-1)$)					MSE ($q = 1/\gamma$)				
λ'	0.0958	0.1740	0.3531	0.9377	26.6313	-27.4316	-1.7225	-1.0514	-0.7921	-0.6252
q	-0.8100	-0.6100	-0.4100	-0.2100	-0.0100	0.0100	0.2100	0.4100	0.6100	0.8100
NMI	0.7539	0.7478	0.7539	0.7478	0.7544	0.7544	0.7070	0.6173	0.6139	0.5485

the learned feature, and $\|\mathbf{W}\|_{1,2} = \sum_{j=1}^{d} \|\mathbf{w}^j\|$ *with* $\mathbf{w}^j$ *being the* j*-th row of* $\mathbf{W}$*. Note that the calibration in above multi-task model is from the use of* ℓ_2 *norm* $\|\mathbf{X}_i\mathbf{w}_i - \mathbf{y}_i\|$ *rather than squared norm* $\|\mathbf{X}_i\mathbf{w}_i - \mathbf{y}_i\|^2$*, which fits into the p-th Root Loss paradigm. Specifically, its Unified Paradigms form* ($\lambda > 0, q = -1$) *is*

$$\min_{\substack{\mathbf{W}\in\mathbb{R}^{d\times c}\\ \alpha\succcurlyeq 0,\alpha^\top \mathbf{1}=1}} \sum_{i=1}^{c} \alpha_i \|\mathbf{X}_i\mathbf{w}_i - \mathbf{y}_i\|^2 + \frac{\lambda}{\alpha_i} + \lambda_1\|\mathbf{W}\|_{1,2} + \frac{\lambda_2}{2}\|\mathbf{W}\|_F^2.$$

5 Numerical Evidence

In this section, we provide numerical evidence for the correctness of our theoretical results. In particular, we conduct experiments on multi-view clustering task and compare the performance of pRL, ED, and UP.

Here we provide the details on the setting of experiments[1]. For the ED and pRL paradigms, we choose the MSE [XTMZ10] (see Example 1) and the AMGL [NLL16] (see Example 3), respectively. In the experiments, we use the MSRCv1 dataset [WJ05] and the experiments setting follows [NTL18]. The commonly used Normalized Mutual Information (NMI) is chosen as the performance measure. All models in our experiments are solved with alternating minimization strategy. Specifically, for the weights learning subproblem, the ED paradigm (MSE) has a closed-form solution (Lemma 6, see appendix); the pRL paradigm (AMGL) is solved with the Iterative Re-weighted strategy (used in [NLL16]); the UP paradigm is solved with the CVX [GB14, GB08] since the subproblem is convex and has very few variables.

The experimental results are reported in Table 2 and 3. We use the columns with $q = \pm 0.0100$ to demonstrate the proposed paradigm UP contains the three weight learning paradigms as special cases. Others columns show that UP can achieve better performance than ED and pRL with properly chosen λ. The λ' of ED/pRL is computed by running ED/pRL first and recording the optimal function objective $\{f_i(x^*)\}_{i=1}^n$. From the proof of Theorem 5, we can see

$$\lambda' = -\mathrm{sgn}(q)\left(\sum_{i=1}^{n}\left(\frac{f_i(x^*)}{|q|}\right)^{1/(q-1)}\right)^{q-1}.$$

We close this section by making several observations from Table 2 and 3:

- UP can reproduce results from NR, ED, and pRL.
- For ED and pRL, the corresponding λ' is not optimal and can be easily improved by tuning λ.
- For almost all γ in ED and p in pRL, the performance can be improved by solving UP with proper λ.
- When using UP in practice, the computed λ' can be used as an initial value for tuning.

6 Conclusion

In this paper, we present a unified paradigm for learning weights to linearly combine multiple views. The unified paradigm contains three widely used weight learning paradigms, *i.e.*, ED, NR, and pRL, as special cases. We provide interesting observations on the setting of the hyper-parameter, the counterintuitive limiting behavior of ED, and present some interesting reformulations. Besides, we believe that our results may inspire further research on new weight learning schemes and be useful in understanding existing algorithms. We conduct numerical simulation and the results support our theoretical analysis well.

Acknowledgments

*Corresponding author: Feiping Nie.
Lai Tian would like to thank Jing Li (now at UTS)

[1] See https://github.com/icety3/unified-paradigm

for helpful conversations on [NLL16], which motivated this work. This work was supported in part by the National Natural Science Foundation of China grant under number 61772427 and 61751202.

References

[BEF84] James C Bezdek, Robert Ehrlich, and William Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203, 1984.

[Ber99] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[BJM+12] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[Bul13] Peter S Bullen. *Handbook of means and their inequalities*, volume 560. Springer Science & Business Media, 2013.

[Chu97] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

[CNCH13] Xiao Cai, Feiping Nie, Weidong Cai, and Heng Huang. Heterogeneous image features integration via multi-modal semi-supervised learning model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1737–1744, 2013.

[CRNA14] Hongmin Cai, Peiying Ruan, Michael Ng, and Tatsuya Akutsu. Feature weight estimation for gene selection: a local hyperlinear learning approach. *BMC bioinformatics*, 15(1):70, 2014.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[GB08] Michael Grant and Stephen Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. http://stanford.edu/~boyd/graph_dcp.html.

[GB14] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

[GMA18] Anil Goyal, Emilie Morvant, and Massih-Reza Amini. Multiview learning of weighted majority vote by bregman divergence minimization. In *International Symposium on Intelligent Data Analysis*, pages 124–136. Springer, 2018.

[GMGA17] Anil Goyal, Emilie Morvant, Pascal Germain, and Massih-Reza Amini. Pac-bayesian analysis for a two-step hierarchical multiview learning approach. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 205–221. Springer, 2017.

[Goy18] Anil Goyal. *Learning a Multiview Weighted Majority Vote Classifier: Using PAC-Bayesian Theory and Boosting*. PhD thesis, Université de Lyon, 2018.

[GZFY14] Pinghua Gong, Jiayu Zhou, Wei Fan, and Jieping Ye. Efficient multi-task feature learning with calibration. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 761–770. ACM, 2014.

[HLP52] Godfrey Harold Hardy, John Edensor Littlewood, and George Pólya. *Inequalities*. Cambridge university press, 1952.

[HYZ+18] Xiaohui Huang, Xiaofei Yang, Junhui Zhao, Liyan Xiong, and Yunming Ye. A new weighting k-means type clustering framework with an l2-norm regularization. *Knowledge-Based Systems*, 2018.

[KM13] Masayuki Karasuyama and Hiroshi Mamitsuka. Multiple graph label propagation by sparse integration. *IEEE transactions on neural networks and learning systems*, 24(12):1999–2012, 2013.

[KR15] Anurag Kumar and Bhiksha Raj. Unsupervised fusion weight learning in multiple classifier systems. *arXiv preprint arXiv:1502.01823*, 2015.

[LJL+15] Jing Liu, Yu Jiang, Zechao Li, Zhi-Hua Zhou, and Hanqing Lu. Partially shared latent factor learning with multiview data. *IEEE transactions on neural networks and learning systems*, 26(6):1233–1246, 2015.

[LJW+17] Hong Liu, Rongrong Ji, Yongjian Wu, Feiyue Huang, and Baochang Zhang.

Cross-modality binary code learning via fusion similarity hashing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7380–7388, 2017.

[LWZ14] Han Liu, Lie Wang, and Tuo Zhao. Multivariate regression with calibration. In *Advances in neural information processing systems*, pages 127–135, 2014.

[NCL17] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, pages 2408–2414, 2017.

[NHL18] Feiping Nie, Zhanxuan Hu, and Xuelong Li. Calibrated multi-task learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2012–2021. ACM, 2018.

[NLL16] Feiping Nie, Jing Li, and Xuelong Li. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *IJCAI*, pages 1881–1887, 2016.

[NTL18] Feiping Nie, Lai Tian, and Xuelong Li. Multiview clustering via adaptively weighted procrustes. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2022–2030. ACM, 2018.

[NWH14] Feiping Nie, Xiaoqian Wang, and Heng Huang. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 977–986. ACM, 2014.

[SSTM17] Shiliang Sun, John Shawe-Taylor, and Liang Mao. Pac-bayes analysis of multi-view learning. *Information Fusion*, 35:117–131, 2017.

[SWF+17] Zhenqiu Shu, Xiaojun Wu, Honghui Fan, Pu Huang, Dong Wu, Cong Hu, and Feiyue Ye. Parameter-less auto-weighted multiple graph regularized nonnegative matrix factorization for data representation. *Knowledge-Based Systems*, 131:105–112, 2017.

[TL10] Grigorios F Tzortzis and Aristidis C Likas. Multiple view clustering using a weighted combination of exemplar-based mixture models. *IEEE Transactions on neural networks*, 21(12):1925–1938, 2010.

[TL12] Grigorios Tzortzis and Aristidis Likas. Kernel-based weighted multi-view clustering. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 675–684. IEEE, 2012.

[WCLC15] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1125–1133, 2015.

[WJ05] J Winn and N Jojic. Locus: learning object classes with unsupervised segmentation. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 756–763. IEEE, 2005.

[XLW+15] Zhe Xue, Guorong Li, Shuhui Wang, Chunjie Zhang, Weigang Zhang, and Qingming Huang. Gomes: A group-aware multi-view fusion approach towards real-world image clustering. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[XNL+17] Xiaowei Xue, Feiping Nie, Zhihui Li, Sen Wang, Xue Li, and Min Yao. A multiview learning framework with a linear computational cost. *IEEE transactions on cybernetics*, 2017.

[XTMZ10] Tian Xia, Dacheng Tao, Tao Mei, and Yongdong Zhang. Multiview spectral embedding. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(6):1438–1446, 2010.

[XTX13] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[XWL16] Yu-Meng Xu, Chang-Dong Wang, and Jian-Huang Lai. Weighted multi-view clustering with feature selection. *Pattern Recognition*, 53:25–35, 2016.

[ZDF17] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *AAAI*, pages 2921–2927, 2017.

[ZF10] Xiao-bin Zhi and Jiu-lun Fan. Some notes on k-harmonic means clustering algorithm. In *Quantitative Logic and Soft Computing 2010*, pages 375–384. Springer, 2010.

[ZHD99] Bin Zhang, Meichun Hsu, and Umeshwar Dayal. K-harmonic means-a data clustering algorithm. *Hewlett-Packard Labs Technical Report HPL-1999-124*, 1999.

[ZHJ+17] Wenzhang Zhuge, Chenping Hou, Yuanyuan Jiao, Jia Yue, Hong Tao, and Dongyun Yi. Robust auto-weighted multi-view subspace clustering with common subspace representation matrix. *PloS one*, 12(5):e0176769, 2017.

[ZHWZ16] Guang-Yu Zhang, Dong Huang, Chang-Dong Wang, and Wei-Shi Zheng. Weighted multi-view on-line competitive clustering. In *Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on*, pages 286–292. IEEE, 2016.

[ZNHY17] Wenzhang Zhuge, Feiping Nie, Chenping Hou, and Dongyun Yi. Unsupervised single and multiple views feature extraction with structured graph. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2347–2359, 2017.

Appendix

A Flowchart of Proofs

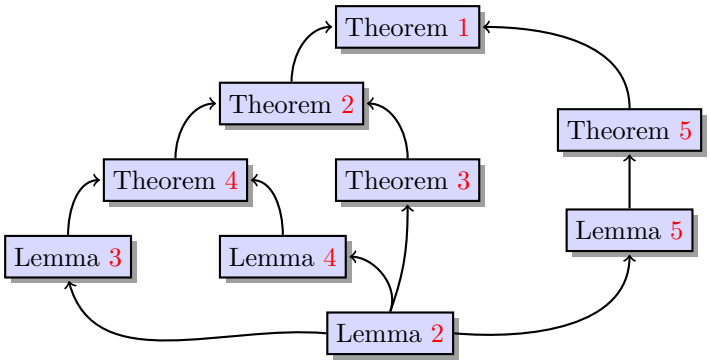

Figure 2: Flowchart of the proofs.

B Deferred Proofs

Lemma 2. *Let*

$$C = \left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} + \lambda\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{q}{q-1}}.$$

Then, we have

$$\min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^q \simeq_x \min_x C \cdot \sum_{i=1}^{n} f_i(x)^{\frac{q}{q-1}}.$$

Proof. Fix x, set the gradient with respect to α to zero and consider the nonnegative constraint. We have

$$\alpha_i^* = \left(\max\left\{\frac{-f_i(x)}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}}.$$

It follows that

$$\begin{aligned}
&\min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^q \\
\simeq_x &\min_x \sum_{i=1}^{n} \left(\left(\max\left\{\frac{-f_i(x)}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} \cdot f_i(x) + \lambda\left(\max\left\{\frac{-f_i(x)}{\lambda q}, 0\right\}\right)^{\frac{q}{q-1}}\right) \\
\simeq_x &\min_x \sum_{i=1}^{n} \left(\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} \cdot f_i(x)^{\frac{q}{q-1}} + \lambda\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{q}{q-1}} \cdot f_i(x)^{\frac{q}{q-1}}\right) \\
\simeq_x &\min_x \left(\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} + \lambda\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{q}{q-1}}\right) \cdot \sum_{i=1}^{n} f_i(x)^{\frac{q}{q-1}},
\end{aligned}$$

which completes the proof. □

Theorem 3. *Given* $0 < p < 1, \frac{1}{p} + \frac{1}{q} = 1$, *then* $q < 0 < \lambda$ *and*

$$\min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^q \simeq_x \min_x \sum_{i=1}^{n} f_i(x)^p.$$

Proof. Note that $q = \frac{p}{p-1} < 0$ since $0 < p < 1$. Thus, we have

$$\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} + \lambda\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{q}{q-1}} = \left(\frac{-1}{\lambda q}\right)^{\frac{1}{q-1}}\left(1 - \frac{1}{q}\right) > 0.$$

Using Lemma 2 and $p = \frac{q}{q-1}$, we have

$$\begin{aligned}
&\min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^q \\
\simeq_x &\min_x \left(\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} + \lambda\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{q}{q-1}}\right) \cdot \sum_{i=1}^{n} f_i(x)^{\frac{q}{q-1}} \\
\simeq_x &\min_x \sum_{i=1}^{n} f_i(x)^{\frac{q}{q-1}} \simeq_x \min_x \sum_{i=1}^{n} f_i(x)^p,
\end{aligned}$$

which completes the proof. □

Lemma 3. *Given $\lambda < 0 < q < 1$, then $\sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^q$ is convex with respect to $\{\alpha_i\}_{i=1}^n$ and*

$$\min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^q \simeq_x \min_x \mathbb{M}_c(\{f_i(x)\}),$$

where $c = \frac{q}{q-1}$.

Proof. Notice that given $\lambda < 0 < q < 1$,

$$\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} + \lambda\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{q}{q-1}} = \left(\frac{-1}{\lambda q}\right)^{\frac{1}{q-1}}\left(1 - \frac{1}{q}\right) < 0.$$

Using Lemma 2 and $\frac{q-1}{q} < 0$, we have

$$\begin{aligned}
&\min_{x,\alpha\succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^q \\
\simeq_x &\min_x \left(\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} + \lambda\left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{q}{q-1}}\right) \cdot \sum_{i=1}^{n} f_i(x)^{\frac{q}{q-1}} \\
\simeq_x &\max_x \sum_{i=1}^{n} f_i(x)^{\frac{q}{q-1}} \simeq_x \min_x \left(\sum_{i=1}^{n} f_i(x)^{\frac{q}{q-1}}\right)^{\frac{q-1}{q}} \simeq_x \min_x \mathbb{M}_c(\{f_i(x)\}),
\end{aligned}$$

which completes the proof. □

Lemma 4. *Let $\gamma > 1$. Then,*

$$\min_{\alpha^\top \mathbf{1}=1,\alpha\succcurlyeq 0} \sum_{i=1}^{n} \alpha_i^\gamma f_i(x) \simeq_x \min_x \mathbb{M}_{\frac{1}{1-\gamma}}(\{f_i(x)\}).$$

Proof. Using the inequality (1) in Lemma 6 and $\frac{1}{\gamma} + \frac{1}{s} = 1$, we have $-\frac{s}{\gamma} = \frac{1}{1-\gamma}$ and

$$\begin{aligned}
&\min_{x,\alpha^\top \mathbf{1}=1,\alpha\succcurlyeq 0} \sum_{i=1}^{n} \alpha_i^\gamma f_i(x) \simeq_x \min_x \left(\sum_{i=1}^{n} f_i(x)^{-\frac{s}{\gamma}}\right)^{-\frac{\gamma}{s}} \\
\simeq_x &\min_x \left(\sum_{i=1}^{n} f_i(x)^{\frac{1}{1-\gamma}}\right)^{1-\gamma} \simeq_x \min_x \mathbb{M}_{\frac{1}{1-\gamma}}(\{f_i(x)\}),
\end{aligned}$$

which completes the proof. □

Theorem 4. *Let $\gamma > 1, q = \frac{1}{\gamma}, \lambda < 0$. Then,*

$$\min_{x,\alpha} \sum_{i=1}^{n} \alpha_i^{\gamma} f_i(x) \simeq_x \min_{x,\alpha} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^{q}.$$

Proof. The strategy is showing that $l.h.s \simeq_x \min_x \mathbb{M}_{c_1}(\{f_i(x)\})$ with Lemma 4 and $r.h.s \simeq_x \min_x \mathbb{M}_{c_2}(\{f_i(x)\})$ with Lemma 3. Then comparing c_1 with c_2 gives the result.

Specifically, given $\gamma > 1$, applying Lemma 4, we have $c_2 = \frac{1}{1-\gamma}$. Meanwhile, we get $\lambda < 0 < q = \frac{1}{\gamma} < 1$ from $\gamma > 1$. Applying Lemma 3, we have $c_1 = \frac{q}{q-1}$. Comparing c_1 and c_2 with $p = \frac{1}{\gamma}$, the result just follows. □

Lemma 5 (Ineffectiveness of $|\lambda|$)**.** *Given $q < 0 < \lambda$ or $\lambda < 0 < q < 1$, we have*

$$\min_{x,\alpha \succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^{q} \simeq_x \min_{x,\alpha \succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + sgn(\lambda) \cdot \alpha_i^{q}.$$

Proof. If $q < 0 < \lambda$ or $\lambda < 0 < q < 1$, then

$$C_1 := \left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} + \lambda \left(\max\left\{\frac{-1}{\lambda q}, 0\right\}\right)^{\frac{q}{q-1}} = \left(\frac{-1}{\lambda q}\right)^{\frac{1}{q-1}} \left(1 - \frac{1}{q}\right)$$
$$C_2 := \left(\left(\max\left\{\frac{-1}{\mathrm{sgn}(\lambda) q}, 0\right\}\right)^{\frac{1}{q-1}} + \lambda \left(\max\left\{\frac{-1}{\mathrm{sgn}(\lambda) q}, 0\right\}\right)^{\frac{q}{q-1}}\right) = \left(\frac{-1}{\mathrm{sgn}(\lambda) q}\right)^{\frac{1}{q-1}} \left(1 - \frac{1}{q}\right).$$

It is easy to see $\mathrm{sgn}(C_1) = \mathrm{sgn}(C_2)$. Thus, applying Lemma 2, we have

$$\min_{x,\alpha \succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^{q} \simeq_x \min_x C_1 \cdot \sum_{i=1}^{n} f_i(x)^{\frac{q}{q-1}} \simeq_x \min_x C_2 \cdot \sum_{i=1}^{n} f_i(x)^{\frac{q}{q-1}} \min_{x,\alpha \succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \mathrm{sgn}(\lambda) \cdot \alpha_i^{q},$$

which completes the proof. □

Theorem 5 (Ineffectiveness of $\alpha^\top \mathbf{1} = 1$)**.** *Given $q < 0 < \lambda$ or $\lambda < 0 < q < 1$, there exists λ' such that*

$$\min_{x,\alpha \succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^{q} \simeq_x \min_{x,\alpha \succcurlyeq 0, \alpha^\top \mathbf{1} = 1} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda' \alpha_i^{q}.$$

Proof. The key idea is that we show the effect of the additional constraint $\alpha^\top \mathbf{1} = 1$ can be cancelled by setting a particular λ' which satisfies $\mathrm{sgn}(\lambda) = \mathrm{sgn}(\lambda')$.

Specifically, denote by x^* the optimal x in $\min_{x,\alpha \succcurlyeq 0} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^{q}$. Also, since $q < 0 < \lambda$ or $\lambda < 0 < q < 1$, let

$$C := \sum_{i=1}^{n} \left(\max\left\{\frac{-f_i(x^*)}{\lambda q}, 0\right\}\right)^{\frac{1}{q-1}} = \sum_{i=1}^{n} \left(\frac{-f_i(x^*)}{\lambda q}\right)^{\frac{1}{q-1}}.$$

Then, consider a new problem

$$\min_{x,\alpha \succcurlyeq 0, \alpha^\top \mathbf{1} = C} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^{q} = \min_x \underbrace{\left(\min_{\alpha \succcurlyeq 0, \alpha^\top \mathbf{1} = C} \sum_{i=1}^{n} \alpha_i f_i(x) + \lambda \alpha_i^{q}\right)}_{g(x)}.$$

Now, for $g(x)$, there will be

1. $x \neq x^* \Rightarrow g(x) \geq g(x^*)$ (since the constraint $\alpha^\top \mathbf{1} = C$ may be active),
2. $x = x^* \Rightarrow g(x) = g(x^*)$ (by definition).

Therefore, if we know C in advance, we can say

$$\min_{x,\alpha\succcurlyeq 0}\sum_{i=1}^{n}\alpha_i f_i(x)+\lambda\alpha_i^q \simeq_x \min_{x,\alpha\succcurlyeq 0,\alpha^\top\mathbf{1}=C}\sum_{i=1}^{n}\alpha_i f_i(x)+\lambda\alpha_i^q.$$

Let $\beta_i=\frac{\alpha_i}{C}, \lambda'=\lambda C^{q-1}$. Then,

$$\min_{x,\alpha\succcurlyeq 0}\sum_{i=1}^{n}\alpha_i f_i(x)+\lambda\alpha_i^q \simeq_x \min_{x,\beta\succcurlyeq 0,\beta^\top\mathbf{1}=1}\sum_{i=1}^{n}\beta_i f_i(x)+\lambda'\beta_i^q,$$

which indicates that the only thing left is to find the oracle C for λ'. A straightforward calculation gives

$$\lambda'=\lambda C^{q-1}=\lambda\left(\sum_{i=1}^{n}\left(\frac{-f_i(x^*)}{\lambda q}\right)^{\frac{1}{q-1}}\right)^{q-1}=-\mathrm{sgn}(q)\left(\sum_{i=1}^{n}\left(\frac{f_i(x^*)}{|q|}\right)^{1/(q-1)}\right)^{q-1}.$$

Using Lemma 5, the proof completes. □

Lemma 6 (Exponential Decay)**.** *Let $\gamma>1$. Then the optimal α^* for problem*

$$\min_{\alpha^\top\mathbf{1}=1,\alpha\succcurlyeq 0}\sum_{i=1}^{n}\alpha_i^\gamma f_i(x) \qquad is \qquad \alpha_i^*=\frac{f_i(x)^{\frac{1}{1-\gamma}}}{\sum_{j=1}^{n}f_j(x)^{\frac{1}{1-\gamma}}}.$$

Proof. The key idea is cancelling the α_i with the Hölder's inequality (Lemma 7).

Note that, for $\gamma>1$ and $\frac{1}{\gamma}+\frac{1}{s}=1$, we have

$$\left(\sum_{i=1}^{n}\alpha_i^\gamma f_i(x)\right)^{\frac{1}{\gamma}}\left(\sum_{i=1}^{n}f_i(x)^{-\frac{s}{\gamma}}\right)^{\frac{1}{s}}\geq\left(\sum_{i=1}^{n}\alpha_i\cdot f_i(x)^{\frac{1}{\gamma}}\cdot f_i(x)^{-\frac{1}{\gamma}}\right)=\left(\sum_{i=1}^{n}\alpha_i\right)=1,$$

which indicates

$$\sum_{i=1}^{n}\alpha_i^\gamma f_i(x)\geq\left(\sum_{i=1}^{n}f_i(x)^{-\frac{s}{\gamma}}\right)^{-\frac{\gamma}{s}}, \tag{1}$$

where the r.h.s is constant with respect to $\{\alpha_i\}_{i=1}^n$.

Thus, when $\alpha_i\propto f_i(x)^{\frac{1}{1-\gamma}}$, the equality holds. Combining with the constraint $\sum_{i=1}^{n}\alpha_i=1$, we reach $\alpha_i^*=f_i(x)^{\frac{1}{1-\gamma}}/(\sum_{j=1}^{n}f_j(x)^{\frac{1}{1-\gamma}})$, which completes the proof. □

Theorem 6. *If hyper-parameters λ and q satisfy*

$$\frac{1}{q}\left(\sum_{i=1}^{k}(x_k-x_i)^{\frac{1}{q-1}}\right)^{q-1}\leq\lambda\leq\frac{1}{q}\left(\sum_{i=1}^{k}(x_{k+1}-x_i)^{\frac{1}{q-1}}\right)^{q-1},$$

then the optimal α^* of above problem has $\|\alpha^*\|_0=k$.

Proof. Note that the objective of NR can be reformulated as

$$\min_{\alpha\succcurlyeq 0,\alpha^\top\mathbf{1}=1}\frac{1}{\gamma}\sum_{i=1}^{n}\alpha_i x_i+\alpha_i^p.$$

The Lagrangian of the reformulated NR can be written as

$$\mathcal{L}(\alpha,\eta,\beta)=\frac{1}{\gamma}\sum_{i=1}^{n}\alpha_i x_i+\alpha_i^p-\eta\left(\sum_{i=1}^{n}\alpha_i-1\right)-\sum_{i=1}^{n}\beta_i\alpha_i.$$

Taking the gradient with respect to α_i, gives

$$\nabla_{\alpha_i}\mathcal{L} = \frac{x_i}{\gamma} + p\alpha_i^{p-1} - \eta - \beta_i.$$

Setting the right-hand side to zero and using the complementary slackness condition, we get α_i with

$$\alpha_i = p^{1/(1-p)} \max(0, \eta - x_i/\gamma)^{1/(p-1)}.$$

Therefore, if we want a k-sparse solution in $\{\alpha_i\}_{i=1}^n$, it should hold

$$\frac{x_k}{\gamma} \leq \eta \leq \frac{x_{k+1}}{\gamma}.$$

Meanwhile, we would say

$$\sum_{i=1}^{k} \alpha_i = \sum_{i=1}^{k} p^{1/(1-p)}(\eta - x_i/\gamma)^{1/(p-1)} = 1.$$

That is

$$\sum_{i=1}^{k} (\eta - x_i/\gamma)^{1/(p-1)} = p^{1/(p-1)}.$$

From the box bound for η, we have

$$\left(\frac{x_k - x_i}{\gamma}\right)^{1/(p-1)} \leq \left(\eta - \frac{x_i}{\gamma}\right)^{1/(p-1)} \leq \left(\frac{x_{k+1} - x_i}{\gamma}\right)^{1/(p-1)}.$$

Summing up with $i = 1 \dots k$, using $\sum_{i=1}^{k}(\eta - x_i/\gamma)^{1/(p-1)} = p^{1/(p-1)}$, we have

$$\sum_{i=1}^{n} \left(\frac{x_k - x_i}{\gamma}\right)^{1/(p-1)} \leq p^{1/(p-1)} \leq \sum_{i=1}^{n} \left(\frac{x_{k+1} - x_i}{\gamma}\right)^{1/(p-1)}.$$

With rearrangement, we obtain

$$\frac{1}{p}\left(\sum_{i=1}^{k}(x_k - x_i)^{\frac{1}{p-1}}\right)^{p-1} \leq \gamma \leq \frac{1}{p}\left(\sum_{i=1}^{k}(x_{k+1} - x_i)^{\frac{1}{p-1}}\right)^{p-1},$$

which completes the proof. □

C Technical Lemmas

Lemma 1 (Power mean inequality [Bul13]). *Given $p < q$, we have*

$$\sqrt[p]{\frac{1}{n}\sum_{i=1}^{n} x_i^p} = \mathbb{M}_p(\{x_i\}) \leq \mathbb{M}_q(\{x_i\}) = \sqrt[q]{\frac{1}{n}\sum_{i=1}^{n} x_i^q}.$$

Specially,

$$\mathbb{M}_{-\infty}(\{x_i\}) = \min_i\{x_i\}, \quad \mathbb{M}_{+\infty}(\{x_i\}) = \max_i\{x_i\},$$

$$\mathbb{M}_0(\{x_i\}) = \prod_{i=1}^{n} \sqrt[n]{x_i}.$$

Lemma 7 (Hölder's inequality [HLP52]). *For $p > 1, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, we have*

$$\sum_{i=1}^{n} |x_i y_i| \leq \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |y_i|^q\right)^{\frac{1}{q}},$$

when there exists $c \neq 0$ such that $|x_i|^p = c \cdot |y_i|^q$ for all $i = 1, \dots, n$, the equality holds.