ON NON-CONVEX SPLITTING METHODS FOR MARKOVIAN INFORMATION THEORETIC REPRESENTATION LEARNING

by

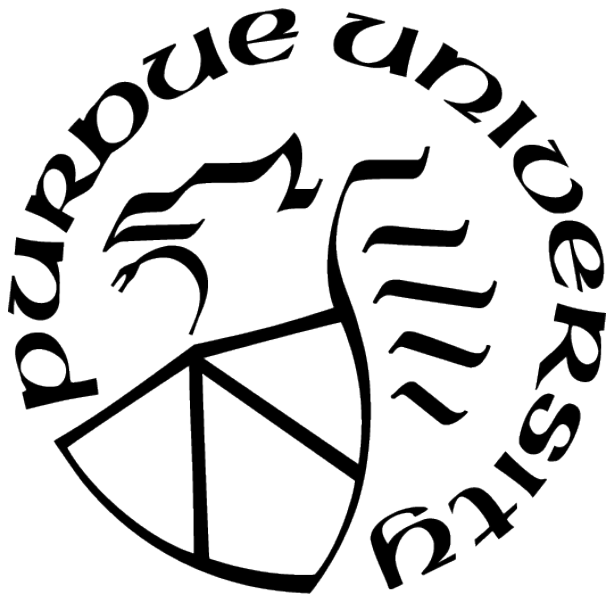
Teng-Hui Huang

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy



School of Electrical and Computer Engineering

West Lafayette, Indiana

May 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Aly El Gamal, Chair

School of Engineering

Dr. James S. Lenert

School of Engineering

Dr. Mark R. Bell

School of Engineering

Dr. Murat Kocaoglu

School of Engineering

Approved by:

Dr. Dimitrios Peroulis

ACKNOWLEDGMENTS

I would like to thank the advisory committee: Dr. Aly El Gamal (chair), Dr. James Lenert, Dr. Mark Bell and Dr. Kocaoglu for their insightful advice, and continuing guidance and support.

I would also like to thank Dr. Hesham El Gamal for inspiring part of the development of this work which enhanced the influence of our results on more challenging but practical problems.

My thanks and appreciations also go to my colleagues: Rehana Mahfuz, Shreya Ghosh, Rohan Manna and Shakti Wadekar. And the people of the Purdue ECE administrative office: Mr. Matt Golden, Ms. Elisheba Van Winkle and Mr. Steven Devault.

Finally, I am very grateful for my family's love and encouragement during the ups and downs of my pursuit for a doctoral degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this work, we study a class of Markovian information theoretic optimization problems motivated by the recent interests in incorporating mutual information as performance metrics which gives evident success in representation learning, feature extraction and clustering problems. In particular, we focus on the information bottleneck (IB) and privacy funnel (PF) methods and their recent multi-view, multi-source generalizations that gain attention because the performance significantly improved with multi-view, multi-source data. Nonetheless, the generalized problems challenge existing IB and PF solves in terms of the complexity and their abilities to tackle large-scale data. To address this, we study both the IB and PF under a unified framework and propose solving it through splitting methods, including renowned algorithms such as alternating directional method of multiplier (ADMM), Peaceman-Rachford splitting (PRS) and Douglas-Rachford splitting (DRS) as special cases. Our convergence analysis and the locally linear rate of convergence results give rise to new splitting method based IB and PF solvers that can be easily generalized to multi-view IB, multi-source PF. We implement the proposed methods with gradient descent and empirically evaluate the new solvers in both synthetic and real-world datasets. Our numerical results demonstrate improved performance over the state-of-the-art approach with significant reduction in complexity. Furthermore, we consider the practical scenario where there is distribution mismatch between training and testing data generating processes under a known bounded divergence constraint. In analyzing the generalization error, we develop new techniques inspired by the input-output mutual information approach and tighten the existing generalization error bounds.

1. INTRODUCTION

Employing information theoretic metrics as objectives in optimizing machine learning models has gained noticeable interests for the remarkable success in supervised and unsupervised classification tasks, feature extraction and representation learning [1]–[4]. Among which, the information bottleneck (IB) [5] and the privacy funnel (PF) [3] methods have drawn substantial attention following the combination with deep neural networks (DNN) for efficient optimization [2]. Perhaps the popularity is a consequence of the formulations that the IB methods aim at finding a latent representation keeping maximum amount of meaningful information while minimize the complexity of expression through processing the observations, and that the PF methods target an optimal mapping with maximum utility at user end whereas minimum leakage of private information [5], [6]. Recently, the IB and PF methods are generalized to learning problems with multi-view, multi-source (multi-modal) data which is expected to give improved performance over single modal data. This aligns with the intuition that one would learn better with text and images given than with either one solely [7]–[13]. However, the attraction comes with new challenges.

In multi-modal regime, the total dimension of data is much larger than the single modal counterpart and existing IB and PF algorithms perform inefficiently or suffer from the "curse of dimensionalty" due to the exponential growth of total number of variables to optimize [14]. Other works resort to heuristic surrogate objectives or rely on black-box DNN architecture to work on multi-modal data that do not fully capture the inherent performance-complexity trade-off with MvIB and MsPF methods. Beyond the empirical success, there is evidently a need for theoretic understanding of the multi-modal representation learning tasks, in vision to better exploit the relevant information within multi-modal data while enjoying significantly reduced complexity. Or conversely, securing privacy within multi-modal data but providing maximally utility to an end-user.

As the goal of representation learning is to achieve reasonable performance when tested with unseen data, the analysis of the generalization error plays an important role for the success of any representation learning approach. While a variety of recent works have adopted IB and PF, or multi-modal generalization, as the objective to form a learning model, its

generalization error analysis is less explored [15], [16]. The recently introduced input-output mutual information-based generalization error bounds provide a new framework to shed light on this direction [17], [18].

1.1 Contributions

Our contributions are summarized as follows. First, we formulate the single-modal IB and PF problems into a unified framework and propose solving them with non-convex splitting methods. The general framework includes a broad class of widely adopted algorithms including alternating directional method of multiplier (ADMM), Peaceman-Rachford splitting (PRS) and Douglas-Rachford splitting (DRS) approaches as special cases. Different from existing ADMM solvers, we exploit the insight that the IB and PF problems can be decoupled into a convex-weak convex pair of sub-objectives, and hence simplifies the three-block design of existing solvers to two-block, which significantly reduces the number of parameters to optimize. Moreover, based on the insight, we develop new IB and PF solvers that can be easily generalized to multi-modal settings. Theoretically, we prove the convergence of the proposed solvers enabled by the recent non-convex ADMM convergence analysis for splitting methods where we further relax some assumptions in literature. Therefore, our convergence analysis is beyond IB and PF problems and applies to a broader class of problems satisfying the convex-weakly convex decoupling. In addition, we prove that the rate of convergence of the proposed approaches are locally linear exploiting the Kurdyka-Łojasiewicz inequality that is recently exploited to characterize the rate of convergence for alternating optimization approaches. The rate is on the same asymptotic order of benchmark solvers for IB and PF known in literature with first-order optimization methods.

Second, we generalize the proposed solvers to multi-modal scenario, the multi-view IB (MvIB) and multi-source PF (MsPF), which extends the splitting algorithms to multi-block, consensus settings. In formulation, we adopt a top-down, information theoretic approach based on the data-processing inequality which is in sharp contrast to existing ones in literature with bottom-up, heuristic design of the overall objective and therefore inherits the relevance-complexity trade-off as in single-modal IB and PF methods, which serves as refer-

ence for performance. We focus on MvIB as the solution easily applies to MsPF with minor modifications. In solving the proposed MvIB objective, we propose two types of solvers catering to the two extremes in terms of the representation overlap. On one end, where there is abundant overlap, we propose a two-step consensus-complement algorithm that first forms a consensus latent representation among multi-view observations then extracts the residual, distinct information that is relevant to the learning task individually given the consensus. On the other end, where there is limited overlap, we propose an incremental-update approach that formulates the latent representation by accumulating the relevant information successively from each view-specific observation. In both cases, we adopt non-convex multi-block consensus ADMM algorithm and prove the convergence and show local linear rate of convergence through the KŁ inequality and further shows that the theoretic convergence guarantee applies to MsPF. Empirically, we implement the proposed MvIB algorithms and compare them to the state-of-the-art deep neural network (DNN) based methods. Remarkably, our results show that the two proposed MvIB algorithms can achieve better performance than the compared method, which demonstrates the advantage of the proposed information-theoretic formulation for representation consensus over the black-box DNN approaches and that our proposed two solvers are endowed with the ability to handle both random and deterministic representation mappings which distinguish them from existing greedy solvers.

Lastly, beyond fitting a known joint probability, we further consider a more general scenario where the goal of the learned representations from a known training distribution is then evaluated with some unknown testing joint distributions. Under this learning theoretic scenario, we propose a new "minimax" framework that captures the distribution mismatch between the learned model and the testing distributions. The framework connects to the recent input-output mutual information generalization error upper bounds. Different from these known results, we propose a novel technique based on the Pythagorean theorem which can be easily applied to most existing bounds and improve the performance. Beyond applications to existing bounds, our theoretical results connect to the recent encoder-decoder structure of learning models in machine learning and provide interpretation to their empirical success. Furthermore, we evaluate our theoretic results on synthetic data and compare

them to existing bounds. The results achieve significant improvements over existing bounds in various supervised learning tasks.

1.2 Literature Review

1.2.1 Information Bottleneck and Privacy Funnel

We start with a review of the IB methods whose development results in the discovery of its dual, the PF problem. Given the joint probability of observations X and a target variable Y, the IB methods aim at finding a representation of raw observations that is minimal in expression complexity but retaining a desired relevance information level toward the latent target. Conventionally, this objective is formulated as the following constrained optimization problem[5]:

$$\begin{aligned} \underset{p(z|x)}{\text{minimize}} \quad & I(X;Z), \\ \text{subject to} \quad & I(Y;Z) > I_0, \\ & \sum_z p(z|x) = 1, \forall x \in \mathcal{X}, \\ & Y - X - Z \quad \text{Markov chain}, \end{aligned} \tag{1.1}$$

where X represents the observations, Y the relevant target and Z the latent representation. The constant threshold $I_0 > 0$ is the desired relevance level and hence controls the trade-off between $I(X;Z)$ and $I(Y;Z)$. In solving the IB problem, a common approach is through the Lagrangian multiplier method, which gives the IB Lagrangian as the minimization objective:

$$\mathcal{L}_{IB} := \gamma I(X;Z) - I(Y;Z) + \sum_x \lambda_x \left(\sum_z p(z|x) - 1 \right), \tag{1.2}$$

where $\{\lambda_i\}_{i\in[|\mathcal{X}|]}$ are multipliers for the equality constraints, imposed to ensure that the conditional probability valid, that is, $p(z|x), \forall x \in \mathcal{X}, z \in \mathcal{Z}$ stays in the compound probability simplex; while the Lagrangian multiplier $0 < \gamma < 1$ is the trade-off parameter controlling the two mutual information $I(Z;X), I(Z;Y)$, known as the relevance-complexity trade-off [6]. This trade-off can be revealed by considering the boundary cases. First, let $\gamma = 0$, then

minimizing the IB Lagrangian is reduces to maximizing a convex function whose optimal solutions are the cases where $p(z|x)$ is equal to 0 or 1. In other words, the deterministic mapping is the optimal solution for $\gamma = 0$. On the other hand, if $\gamma = 1$, then by data-processing inequality (DPI) $I(Z;X) \geq I(Z;Y)$, so the IB Lagrangian is non-negative. Then consider the case where $p(z|x) = 1/|\mathcal{Z}|$, that is, random mapping, the straightforward result $H(Z) = H(Z|X) = H(Z|Y)$ implies $I(Z;X) = I(Z;Y) = 0$. Combine this with DPI, we conclude that the random mapping is the optimal solution for $\gamma = 1$. In fact, this results is recently observed, known as the IB Learnability [19], which turns out to be useful in developing our results in later chapters.

Besides these boundary points, the IB Lagrangian is in general difficult to solve. In literature, only certain types of joint distributions are fully characterized. One well-known example is the Gaussian Information Bottleneck [20] where X, Y are jointly Gaussian distributed. However, for general pairs of variables that might of practical interests, the non-convexity of IB prevents a closed-form solution to fully characterize the relevance-complexity trade-off, i.e. the optimal relevance rate $I(Z;Y)$ one can achieve given a fixed compression rate $I(X;Z)$, without empirically evaluating a given joint probability $p(x,y)$. The non-convexity of the IB problem can be easily revealed as follows. Given a fixed joint probability $p(x,y)$, the mutual information $I(X;Z)$ is a convex function w.r.t. $p(z|x)$; While the mutual information $I(Y;Z)$ is also convex w.r.t $p(z|y)$. From the Markov chain $Y - X - Z$. $p(z|y) = \sum_x p(z|x)p(x|y)$ is a convex combination of $p(z|x)$ and therefore $I(Y;Z)$ is convex w.r.t. $p(z|x)$. Because of the non-convexity, existing algorithms solving the IB problem can assure convergence to local minimizers only. Nonetheless, if the rate of convergence is fast enough, say is at least linearly fast [21], then the price to pay for losing convexity is compensated by multiple trials with random initialization, which is affordable given the advances of computational power today. However, while there have been a variety of algorithms proposed to solve the IB problems, few of them have convergence guarantee.

The most well-known IB algorithm, as first appeared in the seminal work [5], is the Blahut-Arimoto (BA) typed solver [22]. The algorithm is derived through the first or-

der functional derivative of the IB Lagrangian with respect to the conditional probability $p(z|x), \forall z \in \mathcal{Z}, \forall x \in \mathcal{X}$, and iteratively update $p(z|x)$ according to the following steps:

$$\begin{aligned} p^{k+1}(z|x) &= \frac{p^k(z)}{\mathcal{K}(x,\beta)} \exp\Big\{ -\beta D_{KL}[p(y|x)||p^k(y|z)]\Big\}, \\ \mathcal{K}(x,\beta) &= \sum_z p^k(z) \exp\{-\beta D_{KL}[p(y|x)||p^k(y|z)]\}, \\ p^{k+1}(z) &= \sum_{x\in\mathcal{X}} p^{k+1}(z|x)p(x), \\ p^{k+1}(z|y) &= \sum_{x\in\mathcal{X}} p^{k+1}(z|x)p(x|y), \end{aligned} \tag{1.3}$$

where $D_{KL}(\mu||\nu)$ is the Kullback-Leibler (KL) divergence, $\mathcal{K}(x,\beta)$ is the normalization function and the superscript k is the iteration counter. The BA-typed algorithm belongs to a special class of optimization methods, the exponentiated gradient descent [23], which in general is easier to implement as the normalization process automatically project the updated variables to probability simplex. And different from a variety of algorithms inspired by it, the BA-typed solver is non-greedy and has linear rate of convergence assurance [24]. However, due to the general formulation of IB, the later-introduced algorithms, while being greedy or having no theoretic convergence guarantee, have been successfully applied to a variety of problems in various research fields [2], [3], [6], [25], [26]. Among them, the recent combination of deep neural network and IB through variational inference [27] further encourages the application of IB to supervised, unsupervised and reinforcement learning problems [28], [29], which gains even more attention from researchers in various disciplines.

Another kind of algorithm, known as the agglomerative IB [6], limiting the class of conditional probability corresponds to deterministic mapping, further inspires the development of the the privacy funnel (PF) problem [3]. In PF, the variables X, Y now represent public and sensitive information respectively while Z is the observation. This forms the same Markov chain $Y - X - Z$ and the variables to optimize $p(z|x)$ as well. The goal of PF is to find the optimal mapping from a public information to an observation that minimizes the privacy

leakage $I(Y;Z)$ while maximally retaining the utility for the public information $I(X;Z)$. Hence the corresponding PF Lagrangian:

$$\mathcal{L}_{PF} := \beta I(Z;Y) - I(Z;X) + \sum_x \lambda_x \left(\sum_z 1 - p(z|x) \right), \tag{1.4}$$

where the multipliers $\{\lambda_x\}_{x\in\mathcal{X}}$ are applied for $p(z|x)$ to be valid conditional probability; and $\beta > 0$ is the trade-off parameter. In solving (1.4), most existing works adopt the clustering based agglomerative IB approaches. However, due to the limitation on deterministic mappings, this type of solvers is equivalently addressing clustering problems whose complexity scale exponentially as the dimension of X increases. To reduce the difficulty, greedy algorithms are proposed where in each iteration, two clusters are merged if the resultant PF Lagrangian is smaller than treating them as two clusters. While recent works improve the merging process by exploiting the submodularity of the mutual information in clustering scenario [30], [31], the limitation on deterministic mappings remains.

In comparing the performance of existing IB and PF algorithms, a well-known approach for performance evaluation is characterizing the relevance-complexity trade-off for IB, and the privacy-utility trade-off for the PF of the obtained solutions. The evaluation method requires sweeping through a range of the trade-off parameters γ, β and plotting the obtained pairs of $I(X;Z), I(Y;Z)$ in x, y axes. This plot is known as the information plane [5], [6]. By referring an obtained solution of IB or PF to the information plane, one can evaluate the performance. In IB, $I(Y;Z)$ that is higher for a fixed $I(X;Z)$ is considered as a better solution; whereas in PF, for a fixed $I(X;Z)$, a smaller $I(Y;Z)$ is considered as a better solution.

The recent success of applying IB methods to DNN through variational inference and hence allows efficient estimation of the mutual information involved [2]. Interestingly, empirical findings further show the learned representation through the VIB frameworks not only deliver good performance in standard classification accuracy but are also provide robustness to adversarial perturbation without further regularization[2], [32]. This variational IB (VIB [2]) approach has gained significant attention in machine learning and data science community and have been applied to more challenging unsupervised and reinforcement learning

tasks achieving impressive performance. Meanwhile, the advances of machine learning bring new challenges. For example, the ability to tackle large-scale, distributed learning tasks and the privacy awareness in data collection. In these contexts, a class of well-known optimization methods, the alternating direction method of multipliers (ADMM [33]) or more general splitting methods is a promising approach to take the new challenges. Notably, the IB methods have recently been applied to splitting methods which is a timely discovery.

1.2.2 Non-Convex Optimization with Splitting Methods

Recently, [25] adopted ADMM to solve the IB Lagrangian introduced in the last part. While empirically evaluated, there is no convergence guarantee. Inspired by this result, our earlier work simplifies the design to the reference work and proves the convergence of the proposed ADMM-IB solver therein [34]. This theoretic result is based on the recent breakthrough for the convergence analysis of non-convex alternating optimization methods [35].

Compared to the well-studied convex optimization counterpart, non-convex optimization is less explored until recently. The main tool in studying the convergence of non-convex iterative alternating, or the so called non-convex splitting methods, is the Kurdyka-Lojasiewicz (KŁ) inequality [36]–[40]. The KŁ inequality has been successfully applied to the study of non-convex, non-smooth problems and used to characterize the local convergence and the associated rates for a broad class of first-order optimization methods including proximal algorithms, ADMM, Peaceman-Rachford splitting (PRS [41]), Douglas-Rachford splitting (DRS [42], [43]) and even multi-block consensus [44] or sharing algorithms [45].

While in most cases, the splitting methods do not converge if the problem is non-convex, it is recently found that if the non-convexity is composed of a combination of weakly-convex and convex sub-objective functions, then locally linear convergence can be shown through the KŁ inequality under certain smoothness conditions [43], [46], [47].

To apply the KŁ inequality, certain conditions need to be satisfied, known as the KŁ property [38]. The major breakthrough in applying this power tool is that a rich class of functions, functions that are sub-analytic and semi-algebraic, are found to have the KŁ

property [37], [48]. Otherwise, for general function, one needs to find its local convergence behavior through determining of the corresponding Łojasiewicz exponent and then checking each condition needed before applying the KŁ inequality. One of the benefits from applying KŁ inequality is the characterization of the rate of convergence with respect to the associated Łojasiewicz exponent. While explicitly determining the exponent is not a straightforward task [49], once determined, one instantly knows the rate of convergence. For example, if the exponent $\theta = 1/2$ then it corresponds to linear rate. Another benefit is that the convergence characterized by the KŁ inequality is point-wise. Therefore, if the rate of convergence is linear, indicated by the KŁ inequality, then the convergence is Q-linear, a stronger sense of linear convergence compared to R-linear rates [21].

This recent development of the fundamentals for non-convex, non-smooth optimization meets the shift of interests toward large-scale data analysis and learning problems. As it is well-known that, in convex settings, splitting methods decouple the objective functions, allowing efficient parallel computing, the recent results on non-convex settings can bring this computational gain to broader class of practical problems. Among which, the multi-view IB and multi-source PF are perfect candidates to apply the non-convex splitting methods.

1.2.3 Multi-Modal Representation Learning

Owing to the advances in data science and machine learning, there is a recent surge of interests in learning from multi-view data in the machine learning and data science community (e.g., [11], [50]–[53]). What does a view mean? We define a view as a description or observation about the source of data. For example, a news article can be written in different languages, a video can be either colorful or gray scale and wireless signals can be represented in either time or frequency domains. Intuitively, one will expect learning from multi-view data can result in improved performance and several empirical evidence has been found to encourage this approach by recent works [54]. While there is a recent surge of interests on multi-view learning, the study of learning problems began quite early. In literature, the approaches in multi-view clustering and learning can be categorized into two groups: The first group of methods treats the features of all views, forming a single giant view whereas the

second group of paradigms extracts the relevant features of each view, forming a consensus of representations as a low dimensional latent feature. The pros and cons of each group of approaches are clear: combining all view of features into a single one keeps all the information that each view has but the dimension of the giant view increases exponentially, which makes it prohibitively complex to tackle; Whereas the formation of lower-dimensional consensus features significantly reduces the complexity but finding the criterion and developing a general method to construct consensus among all views are challenging tasks.

For the latter group, in search of a general criterion for the formation of consensus, several works have attempted to generalize the single-view IB methods to multi-view IB (MvIB) representation learning. In particular, exploiting the compression and regularization effect by imposing the mutual information of observations and the learned representation, [8] employed mutual information of view-specific observations/representations to the main objective function. Their goal was to maximize the marginal of a linear classifier and the focus was on binary tasks. By treating each view as a single-view IB problem followed by a post-processing stage, [55] proposed maximizing the mutual information between the pairs of view-specific representations to facilitate the formation of consensus among view-specific clustering hypotheses. More recently, [7] proposed a bottom-up, heuristic MvIB objective which is composed of a combination of view-specific information, view-shared information, and inter-cluster correlation measured in mutual information between each information and the latent feature. They propose maximizing the combination of the sub-objectives expressed in mutual information while compressing the observations simultaneously. In minimizing the overall objective function, the reference work resorts to using a class of off-the-shelf greedy single-view IB algorithms, as reviewed in Chapter 1.2.1, and select the best result among them. In MvIB, the state-of-the-art method applies VIB to address the difficulty in forming a consensus for all views. In [9] each view was pre-processed in single-view IB fashion and then each individual-view representation was merged by an additional layer of neurons whose output can be interpreted as the latent consensus of all representations, with *heuristic* loss function imposed on the output aiming to maximize the relevance of the combined representation to a target variable.

On the other hand, in accordance with the duality of IB and PF, the dual problem of MvIB, the multi-source PF (MsPF), where the private information in the original PF formulation now generalized to a class of sensitive attributes [10], [12], [13]. Then, motivated by the advantage in MvIB that prevents the exponential scaling of merging all attributes into one, the goal is to find a mapping scheme from the public information to the observation at the user end that minimize the privacy leakage to each sensitive attribute while maximally maintains the utility of the observation for a user. The MsPF problem is less explored and the existing solvers again resort to using clustering-based greedy algorithms which are limited to deterministic mappings. Aiming at improving the exploration of the privacy-utility trade-off in this generalized MsPF problem, we again apply the recent convergence non-convex splitting methods to develop new class of algorithms that can tackle both random and deterministic mappings and more importantly, provide convergence guarantee. For convenience, in the following chapters, we will term the MvIB and MsPF together as the multi-modal representation learning problems.

1.2.4 Information Theoretic Generalization Error Bounds

Representation learning, either with single or multi modals tasks, is a topic of learning theory. The fundamental question for learning is how well a model performs when tested with unseen samples, after trained with a finite number of available data. In other words, the goal is hoping a model to acquire the informative features from history that allows it to infer the future. In literature, the most popular approach in realizing this objective is known as the empirical risk minimization (ERM [14]), where a certain surrogate loss function is imposed on a learning model, then a set of algorithms is adopted to minimize the empirical estimate of the loss with a finite number of training data. However, as mentioned earlier, the expected value of the loss with the true data-generating distribution is what the learner care most. The absolute difference between the two estimated losses is defined as the generalization error. Intuitively, if the expected loss of ERM is close to that with the unknown data-generating distribution then one can claim that the model learns well. Nonetheless, this depends on a

judicious choice of the loss function and enough training data which is infeasible in general learning tasks.

Finding a tight upper bound of the generalization error has been one of the main topics in learning theory. Conventionally, the approaches adopted to study the generalization error require defining certain complexity metrics that can be crudely categorized into two ideas. The first one is based on the intuition that the cause of generalization error is due to the finite sampling process, and therefore the generalization error can be reduces as the number of training data samples increases for a finite hypothesis space, i.e. the collection of learning models that one algorithm can produce. Some celebrated approaches include the Rademacher Complexity and Vapnik-Chervonenkis (VC) dimension [14], [56]. The second idea instead considers the stability of a learning algorithm, that a stable algorithm should intuitively produce similar outcomes if the training data is replaced by a single sample. Based on this intuition, the uniform stability [57] put more focuses on the learning algorithm where certain regularization on the hypothesis space and Lipschitz smoothness conditions are imposed on the loss functions and hence take parts in the resultant generalization error upper bounds. Recently, the so called input-output mutual information based generalization error bounds approaches are introduced that can incorporate both the sample complexity and the algorithmic stability into an information-theoretic framework for the generalization error analysis [17], [18], [58]–[61]. Through the combination of the sub-Gaussian the assumption along with the Donsker-Varadhan representation of the Kullback-Leibler (KL) divergence [62]. This information-theoretic approach is recently found to achieve state-of-the-art tight upper bounds owing to a variety of upper bound tightening techniques following its emergence [59]–[61], [63], and hence gains significant attention fueled by the recent interests in optimizing information-theoretic metrics as surrogate loss functions [64].

Most existing approaches, however, do not take the adversarial scenario into account. The recent discovery of the existence of adversarial samples create new challenges to generalization error analysis. In adversary learning scenario, it is assumed there is an attacker who craftily adding certain metric-bounded perturbation to clean samples, but dramatically deteriorate the performance of a learning model that performs reasonably well on clean data instead. As the concern for security and privacy draw significant attention along with the

prevalent success of DNN in machine learning, one will also be interested in studying generalization error bounds with the presence of an adversary. Since the discovery of adversarial samples, new defense and attack approaches have been introduced but there is no telling who triumph so far. This everlasting competition further urges the need of theoretic analysis of generalization error bounds.

In an attempt to address the above-mentioned challenges. We view the generalization error as distribution mismatch that might be caused from either sampling process of an unknown data-generating process or a metric-bounded adversary [58], [61]. Different from existing works that also consider distribution mismatch, we formulate the generalization error problem into a minimax problem. In solving the proposed minimax problem, we derive surrogate upper bound for the inner maximization problem that extends the mutual information based upper bounds to cases where indirect partial knowledge of test distribution are available, which is motivated by the fact that adversarial samples need to be plausible to the clean data and hence fall within some bounded-divergence to training distribution, providing additional information to exploit. Moreover, even without the partial knowledge, we find that the derived bounds can tighten the existing mutual information based bounds and apply to a variety of upper-bound tightening techniques. Notably, when imposed the proposed framework on the recently popular encoder-decoder structural learning models, we find that our results connect to the strong data-processing inequality can potentially provide explanations to their empirical success that is less addressed in literature.

2. MARKOVIAN INFORMATION THEORETIC OPTIMIZATION

2.1 Problem Formulation

2.1.1 A General Framework

In this chapter, we study the following Markovian information theoretic Lagrangian:

$$\mathcal{L} := \rho_z H(Z) + \rho_{z|x} H(Z|X) + \rho_{z|y} H(Z|Y), \tag{2.1}$$

where $\rho_z, \rho_{z|x}, \rho_{z|y}$ are some constants to be decided, $H(X)$ denotes the Shannon entropy function of a random variable X whereas $H(X|Y)$ denotes the conditional entropy function of X conditioned on Y. We focus on discrete random variables and leave the generalization to continuous settings to future exploration. Note that since the negative (conditional) entropy function is convex [65], if the coefficient $\rho < 0, \rho \in \{\rho_z, \rho_{z|x}, \rho_{z|y}\}$, then is becomes a convex function. To further simplify (2.1), we define the following vectors.

$$p_z := \begin{bmatrix} p(z_1) & \cdots & p(z_K) \end{bmatrix}^T, \tag{2.2a}$$

$$p_{z|x} := \begin{bmatrix} p(z_1|x_1) & \cdots & p(z_1|x_N) & p(z_2|x_1) & \cdots & p(z_K|x_N) \end{bmatrix}^T, \tag{2.2b}$$

$$p_{z|y} := \begin{bmatrix} p(z_1|y_1) & \cdots & p(z_1|y_M) & p(z_2|y_1) & \cdots & p(z_K|y_M) \end{bmatrix}^T, \tag{2.2c}$$

where the cardinalities of the variables are denoted as $K := |\mathcal{Z}|, N := |\mathcal{X}|, M := |\mathcal{Y}|$ respectively. From the above definition, $p_{z|x}, p_{z|y}$ are cascaded conditional probabilities in vector form. Similarly, we have the vector forms of the prior probabilities p_x, p_y. Moreover, if we arrange the conditional prior probability $p(x|y)$ into a matrix $W_{x|y}$ with the (i, j)-entry $p(x_\mathrm{i}|y_\mathrm{j})$, then under the Markov chain $Y - X - Z$, we can express the relation as a linear equation $p_{z|y} = Q_{x|y} p_{z|x}$, and the marginal relation $p_z = Q_x p_{z|x}$ where $Q_{x|y} := I_K \otimes W_{x|y}^T, Q_x := I_K \otimes p_x^T$ with the operator $\otimes$ denotes the Kronecker product. With these definitions, the (conditional) entropy functions can be expressed as real value functions that take vectors as inputs. Then

we propose solving (2.1) with two-block splitting methods, which gives the following augmented Lagrangian:

$$\mathcal{L}_c(p, q, \nu) := F(p) + G(q) + \langle \nu, Ap - Bq \rangle + \frac{c}{2}\|Ap - Bq\|^2, \tag{2.3}$$

where p, q are primal variables which can be one of $p_z, p_{z|x}, p_{z|y}$ or a cascaded long vector with two of them. The design of p, q subsequently determines the functions F, G; ν denotes the dual variables in vector form to approximate the Lagrange multipliers imposed on the linear constraint $Ap - Bq$ composed of the matrices A, B and primal variables. Since p, q are probability vectors, the meaning of the linear constraint is simply the Markov or marginal probability relation. Lastly, the penalty coefficient $c > 0$ is introduced along with the squared 2-norm of the linear constraint, also known as the proximal term. This penalty term is imposed on the augmented Lagrangian to encourage the constraint to be satisfied which is in sharp contrast of standard Lagrange multiplier methods.

2.1.2 Examples: Information Bottleneck and Privacy Funnel

The proposed general Markovian Lagrangian (2.1) includes a broad range of problems, that is, a different choice of the the coefficients $\rho_z, \rho_{z|x}, \rho_{z|y}$ corresponds to a distinct problem. Among which, we are particularly interested in the assignments of $(\rho_z, \rho_{z|x}, \rho_{z|y})$ such that the overall Lagrangian corresponds to non-convex optimization problem. In all the non-convex assignments, we show in the following that the IB and PF problems fall within this non-convex subset. For the IB methods, by selecting the coefficients as:

$$\rho_z := \gamma - 1, \quad \rho_{z|x} := -\gamma, \quad \rho_{z|y} := 1,$$

whose corresponding Lagrangian is:

$$\mathcal{L}_{IB} := (\gamma - 1)\, H(Z) - \gamma H(Z|X) + H(Z|Y) = \gamma I(Z; X) - I(Z; Y).$$

Note that $\gamma \in (0, 1]$, which is due to the recent result that shows that $\gamma \geq 1$ corresponds to trivial solutions as shown in Chapter 1.2.1. On the other hand, as for the PF problem, we can instead select the coefficients as:

$$\rho_z := \beta - 1, \quad \rho_{z|x} := 1, \quad \rho_{z|y} := -\beta,$$

so that the Lagrangian is:

$$\mathcal{L}_{PF} := (\beta - 1)\, H(Z) + H(Z|X) - \beta H(Z|Y) = \beta I(Z;Y) - I(Z;X),$$

where $\beta > 0$. As discussed in Chapter 1.2.1, the IB and PF problems have attracted significant attention recently due to their success in machine learning and data science research, and new challenges that come along with them. Hence, instead of finding more examples that belong to the proposed generalized framework, in the rest of this chapter, we will focus on the two problems and present a novel optimization-mathematics perspective of the IB and PF problems.

2.2 Algorithms

In solving the augmented Lagrangian (2.3), we consider two types of iterative algorithms. Both algorithms correspond to two-block splitting methods [46]. Note that this simplifies the design and enjoys better convergence assurance than other existing splitting methods based IB three-block solver [25], [34].

2.2.1 First Kind: F-Dual Splitting Algorithm

Denote the superscript k as the iteration counter. The update of the primal and dual variables from step k to step $k+1$ follows:

$$\begin{aligned}
\nu_{1/2}^{k+1} &:= \quad \nu^k - (1-\alpha)c(Ap^k - Bq^k), && \text{(2.4a)}\\
p^{k+1} &:= \quad \underset{p\in\Omega_p}{\arg\min} \quad \mathcal{L}_c(p, q^k, \nu_{1/2}^{k+1}), && \text{(2.4b)}\\
\nu^{k+1} &:= \quad \nu_{1/2}^{k+1} + c(Ap^{k+1} - Bq^k), && \text{(2.4c)}\\
q^{k+1} &:= \quad \underset{q\in\Omega_q}{\arg\min} \quad \mathcal{L}_c(p^{k+1}, q, \nu^{k+1}), && \text{(2.4d)}
\end{aligned}$$

where $\nu_{1/2}^k$ denotes the relaxation step with the corresponding relaxation coefficient $\alpha > 0$. We stress the difference of $\nu_{1/2}^k$ to ν^k where the latter specifically indicates the dual variable at step k.

2.2.2 Second Kind: G-Dual Splitting Algorithm

Alternatively, we have the second type of iterative algorithm that updates the primal and dual variables from step k to step $k+1$ according to the following iterative algorithm:

$$\begin{aligned}
p^{k+1} &:= \quad \underset{p\in\Omega_p}{\arg\min} \quad \mathcal{L}_c(p, q^k, \nu^k), && \text{(2.5a)}\\
\nu_{1/2}^{k+1} &:= \quad \nu^k - (1-\alpha)c(Ap^k - Bq^k), && \text{(2.5b)}\\
q^{k+1} &:= \quad \underset{q\in\Omega_q}{\arg\min} \quad \mathcal{L}_c(p^{k+1}, q, \nu^{k+1}), && \text{(2.5c)}\\
\nu^{k+1} &:= \quad \nu_{1/2}^{k+1} + c(Ap^{k+1} - Bq^k). && \text{(2.5d)}
\end{aligned}$$

The difference to the first algorithm is that the primal variables p, q are updated preceding the relaxation step $\nu_{1/2}$ and the dual ascend of ν. As will be shown in the convergence analysis, if assume smoothness on the sub-objective functions F, G and solved with first-order optimization methods (gradient descent), the first algorithm has its dual variable connects to the smoothness of F while the second links its dual variable to the smoothness

of G instead. We exploit these connections to derive the convergence and the associated rates.

2.3 Main Results

Our main results for this chapter are the theoretic convergence analysis (Chapter 2.4) of the two algorithms proposed to solve the general Markovian information-theoretic optimization problem (4.20). Our optimization-mathematics approach in studying the general problem allows us to prove not only the convergence but also the linear rate of convergence of the algorithms in non-convex settings which is in sharp contrast to the well-studied convex counterparts. We summarize the results in Table 2.1 which serves as a guidance in designing algorithms to solve non-convex information theoretic optimization problems that fall within the proposed framework (2.1). We apply the results to design new solvers for two special cases of the proposed framework, the IB and PF problems (Chapter 2.5) that recently gain significant research interests for its success in machine learning and data science studies. Compared to existing algorithms in each problem, we prove that the proposed new IB solvers achieve asymptotic convergence rate of benchmark solvers (e.g. Blahut-Arimoto typed [5], [22]). Furthermore, in contrast to existing ADMM solvers [34] for IB, we simplify the design where no additional regularization terms are imposed on the augmented Lagrangian, and improves the performance in terms of the smallest penalty coefficient that assures convergence. As for PF, our new DRS solver can tackle both random and deterministic mappings and is therefore different from existing greedy algorithms that are limited to deterministic mappings only. Empirically, we evaluate the new solvers (Chapter 2.6) on both synthetic and real-world datasets and compare the performance to a range of existing solvers. The results demonstrate that the new IB solvers achieve comparable performance in characterizing the relevance-complexity trade-off [6], and that the new PF solvers can better characterize the privacy-utility trade-off than existing algorithms.

Table 2.1. Summary of the Convergence Analysis for Two-Block Non-convex Splitting Methods

Reference	**Algorithm**	**Convergence Conditions**	**Rate of Conv.**	**Properties of Functions**	**Linear Constraints** $Ap - Bq$
Algorithm (2.4)	$0 < \alpha \leq 2$	$c > \max\{\omega_G, \frac{L_p+\sigma_F}{\mu_A^2\alpha}\}$	locally linear	F:σ_F-strongly convex G:ω_G-restricted weakly cvx. F is L_p-smooth	A:positive definite
Algorithm (2.5)	$0 < \alpha < 2$	$c > \frac{\alpha\sigma_G+\pi_q}{(4-2\alpha)\mu_B^2}$ *	locally linear	F:convex G:σ_G-weakly convex G is L_q-smooth	A:full row rank B:positive definite
Algorithm (2.5)	$0 < \alpha < 2$	$c > M_q[\frac{\alpha\sigma_G M_q+\phi_q}{4-2\alpha}]$ †	locally linear	F:convex G:σ_G-weakly convex G:M_q- Lipschitz continuous G are L_q-smooth	A:positive definite, B:full row rank
Jia et al. [47]	Prox. ADMM $(\alpha = 1)$	$c > \frac{\sigma_G+\sqrt{\sigma_G^2+8L_q^2}}{2\mu_B^2}$	locally linear	F:convex G:σ_G-weakly convex G is L_q-smooth	A:positive definite B:positive definite
Themelis et al. [43]	$0 < \alpha < 2$; $2 \leq \alpha < 4$	$c > L_p$; $\frac{\alpha/L_p-\delta_p}{4} < c < \frac{\alpha/L_p+\delta_p}{4}$ ‡	best case sublinear $\mathcal{O}(1/\sqrt{k})$	F:σ_F-hypo-cvx. $(0 < \alpha < 2)$; F:σ_F-str. cvx. $(2 \leq \alpha < 4)$ F is L_p-smooth	$A = B = I$

* $\pi_q := \sqrt{\alpha^2\sigma_G^2 + 8(2-\alpha)L_q^2}$, † $\phi_q := \sqrt{M_q^2\sigma_G^2\alpha^2 + 8(2-\alpha)\lambda_B^2/\mu_{BB^T}}$, ‡ $\delta_p := \sqrt{(\sigma_F\alpha/L_p)^2 - 8(\alpha-2)\sigma_F/L_p}$.

2.3.1 Formulating to Two-Block Augmented Lagrangian

We propose two ways to formulate the three-block Markovian Lagrangian (2.1) into two-block augmented Lagrangian (2.3). The first technique is **linear composition** where we impose $p_{z|y} = Q_{x|y}p_{z|x}$ to be strict equality, that is, the Markov relation between the two probability vectors defined in (2.2). By linear composition, the variable $p_{z|y}$ reduces to a linear transform of $p_{z|x}$. Then the sub-objective function $\rho_{z|y}H(Z|Y)$ can then be expressed as a function of $p_{z|x}$. It turns out that for IB, imposing $p_{z|y} = Q_{x|y}p_{z|x}$ recovers our earlier work [34] while in PF, imposing $p_z = Q_x p_{z|x}$ introduces a new class of non-greedy, linearly convergent, splitting methods based solver. The second technique is **stacking** where two of the three probability vectors involved in the Markovian Lagrangian are cascaded and expressed as a giant vector. For example, by stacking $q := \begin{bmatrix} p_z^T & p_{z|y}^T \end{bmatrix}^T$, the equivalent sub-objective function $G(q) := \rho_z H(Z) + \rho_{z|y}H(Z|Y)$ is then expressed as a function of q. In section 2.5, we show that by stacking, we can introduce a new type of splitting methods-based IB solver that can easily be generalized to multi-view IB representation learning problems (Chapter 3) while the other type cannot. Note that after applying any of the two techniques, the associated sub-objective function may lose convexity/concavity in its original form. For example, in IB, applying the linear composition $p_{z|y} = Q_{x|y}p_{z|x}$ makes $G(p_{z|x}) = -\gamma H(Z|X) + H(Z|Y)$ a non-convex function w.r.t. $p_{z|x}$. As a final remark, recall that the goal of solving the Markovian Lagrangian is to find the mapping $p_{z|x}$, so after "over-parameterizing" to adopt the splitting methods, suppose the two-block splitting algorithms converge, then the next step is to decompose the component $p_{z|x}^*$ that solves both the augmented and the Markovian Lagrangians as the solution. Because of this, the matrices A, B in the linear penalty $Ap - Bq$ that is applied to $p_{z|x}$ need to be at least full column rank.

2.3.2 Linearly Convergent Splitting Methods-Based Solvers

Besides introducing new formulations, we provide theoretical convergence analysis of the proposed methods under three sets of assumptions that apply to two types of IB and one type of PF solvers. The analysis closely follows the recent convergence results for non-convex non-smooth alternating algorithms in first-order optimization methods [39], [45], [66]. Among a

variety of formulations and assumptions considered in this line of research, we found that the "convex-weakly convex" pair sub-objective functions setup is the closest one to ours [46], [47]. As the name of this class suggested, the non-convexity of the original objective function lies in the weakly-convex part of the sub-objective function. It turns out that if the weakly convex sub-objective is also Lipschitz smooth, then convergence can be proven by combining the smoothness and the first-order optimal conditions of the updates for the variables ν and $\nu_{1/2}$. Moreover, the recent results also extend to the rate of convergence analysis which is based on the Kurdyka-Łojasiewicz (KŁ) inequality [36]–[38]. The KŁ inequality (Definition 2.4.7) characterizes the rate of convergence around the neighborhood of a stationary point in terms of the associated Łojasiewicz exponent $\theta \in (0, 1)$ of a function f. While determining the exponent is a difficult task in general [49], there is a broad class of functions that is recently found to satisfy the KŁ inequality and whose exponents are known [48]. Unfortunately, in our case, due to the reformulation to the two-block augmented Lagrangian, we need to explicitly find the Łojasiewicz exponent $\theta = 1/2$ (corresponds to the linear rate of convergence) explicitly.

Following the discussion, the key element to connect to these recent results depends on the smoothness of the sub-objective functions. However, in our case, the functions are either (negative) entropy or conditional entropy which do not have uniform Lipschitz smoothness as the gradients at the boundary points of the probability simplex are ill-defined. Inspired by the smoothness assumptions in density and entropy estimation research, we define ε-infimality (Definition 2.4.2), which avoids these undesired special cases. When imposed on the probability vectors, then the smallest entry of each of the vectors is bounded away from zero by a positive constant ε. In other words, the condition regularizes the feasible sets when solving the augmented Lagrangian. Interestingly, the ε-infimality conditions are realized when implementing the step-size selection algorithm for first-order methods. This implies that the conditions are more than theoretical simplifications but also involve practical aspects of the proposed methods, but this is beyond the scope of this work and will be left for future explorations.

In the next section, we start with the most restricted set of assumptions, where the two sub-objectives satisfy strong convexity and restricted weak convexity (Definition 2.4.4) and

both sub-objective functions are smooth. The convergence can be shown by establishing the sufficient decrease lemma (Lemma 2.4.6). And then by explicitly showing that the Łojasiewicz exponent $\theta = 1/2$ (Lemma 2.4.14), we can adopt the KŁ inequality (Lemma 2.4.10) to prove locally linear rate of convergence (Theorem 2.4.15). We apply this result to develop the first type of DRS-based IB solver (Theorem 2.5.1).

To relax the first set of assumptions, we consider the objective function to be a "convex-weakly convex" pair. Interestingly, now we only need to impose smoothness conditions on the weakly convex part to prove convergence (Lemma 2.4.8) and rate of convergence (Theorem 2.4.18). Based on this result, we develop the second type of DRS-based IB solver (Theorem 2.5.2).

Lastly, for PF, after formulating to the augmented Lagrangian following the linear composition $p_z = Q_x p_{z|x}$, it reveals that the two sub-objective is also the "convex-weakly convex" pair. However, the rank conditions are in consistent with the second set of assumptions. To address this, following the key result in [45], the M-Lipschitz continuity of the weakly convex function provides the reverse control of the non-trivial relation $\|p_{z|x}^k - p_{z|x}^{k+1}\| \leq M\|Q_x p_{z|x}^k - Q_x p_{z|x}^{k+1}\|$ and hence we arrive at the locally linear rate of convergence (Theorem 2.4.20). This result gives rise to a new class of non-greedy PF solvers (Theorem 2.5.4).

2.4 Convergence Analysis

In this section, we present the theoretic convergence analysis of the two algorithms (2.4) and (2.5) under three sets of assumptions. As will be shown in next section, the three different sets of assumptions correspond to two types of IB solvers and a class of PF solvers. Continuing the proof, for each set of assumption, we follow the recent results for the convergence analysis on non-convex the splitting methods but extend from ADMM to the more general DRS through incorporating an additional relaxation step [43].

2.4.1 Preliminaries

Definition 2.4.1. *A function $f : \mathbb{R}^d \mapsto [0, \infty)$, with distinct $x, y \in \Omega$ is Lipschitz continuous if:*

$$|f(x) - f(y)| \leq M|x - y|.$$

Where $M > 0$ is the Lipschitz coefficient.

Note that if $f \in \mathcal{C}^d, d \geq 1$ and $\nabla f(x)$ is *L-Lipschitz continuous*, then the function f is said to be a *L-smooth* function. To avoid confusion, we will denote the Lipschitz continuity coefficient as M whereas the coefficient for smoothness is L.

Definition 2.4.2. *A measure $u(x) \in (\mathcal{X}, \mathcal{F})$ is said to be ϵ-infimal if there exists $\epsilon > 0$, such that $\inf_{x \in \mathcal{X}} u(x) = \varepsilon$.*

The infimal measure condition is commonly assumed in non-parametric entropy/density estimation to assure smoothness on the estimators and hence facilitate the optimization process [67], [68]. A ϵ-infimal measure has its smallest mass strictly bounded away from zero by a positive constant ϵ. As it turns out, the convergence of non-convex information-theoretic optimization problems relies on this condition not only for theoretic analysis, but also in practical application as the infimal coefficient ε translates to the step-size selection algorithm when implemented with first-order gradient descent methods.

Lemma 2.4.1. *let $f(u) = \sum_{\mathrm{i}=1}^{|\mu|} \mu_{\mathrm{i}} \log \mu_{\mathrm{i}}, \sum_{x \in \mathcal{X}} \mu(x) = 1$ be the negative entropy function where two distinct measures $\mu, \nu \in (\mathcal{X}, \mathcal{F})$ are ϵ-infimal. Then f is $|\log \epsilon|$-Lipschitz continuous and $1/\epsilon$-smooth*

Proof. The Lipschitz continuity follows:

$$\begin{aligned} f(\mu) - f(\nu) &= \sum_x [\mu(x) - \nu(x)] \log \frac{1}{\nu(x)} - D_{KL}(\mu||\nu) \\ &\leq (\log \frac{1}{\epsilon}) \sum_x |\mu(x) - \nu(x)| \\ &= |\log \epsilon| \|\mu - \nu\|. \end{aligned}$$

As for the smoothness condition, we have:

$$|\nabla f(\mu) - \nabla f(\nu)| \leq \frac{|\mu - \nu|}{\min_{x\in\mathcal{X}}\{\mu(x), \nu(x)\}} = \frac{|\mu - \nu|}{\epsilon}, \tag{2.6}$$

where the inequality is due to the following identity and the fact that $\log x < x - 1$ for $x > 0$:

$$\begin{cases} a > b, & \log \frac{a}{b} \leq \frac{a}{b} - 1 = \frac{a-b}{b} \\ b > a, & \log \frac{b}{a} \leq \frac{b}{a} - 1 = \frac{b-a}{a} \end{cases} \Rightarrow \left|\log \frac{a}{b}\right| \leq \frac{|a-b|}{\min\{a, b\}}.$$

□

Similarly, for conditional entropy, if assuming infimality on the associated conditional probability vector, then we can establish the following smoothness condition.

Corollary 2.4.2. *Let p_x be given, $p_{z|x}$ be ϵ-infimal, then the conditional entropy $H(Z|X) = -\sum_x p(x) \sum_z p(z|x) \log p(z|x)$ is $|\log \epsilon|$-Lipschitz continuous and $1/\epsilon$-smooth.*

Proof. Following lemma 2.4.1, for two measures $u, v \in \Omega_{z|x}$, where $\Omega_{z|x}$ denotes a compound simplex for the conditional probability $p(z|x)$, the Lipschitz continuity follows:

$$\begin{aligned} H(Z^m|X) - H(Z^n|X) &\leq |\log \epsilon| \sum_x p(x) \sum_z |p(z^m|x) - p(z^n|x)| \\ &\leq |\log \epsilon| \sup_{x\in\mathcal{X}} p(x) \|p^m_{z|x} - p^n_{z|x}\| \\ &= |\log \epsilon| \, \|p^m_{z|x} - p^n_{z|x}\|. \end{aligned}$$

On the other hand, to prove the smoothness, like the right-hand side of the inequality (2.6), we have:

$$|\nabla H(u) - \nabla H(v)| \leq \frac{\max_{x\in\mathcal{X}} p(x)}{\epsilon} |u - v| \leq \frac{|u - v|}{\epsilon}.$$

□

Definition 2.4.3. *A differentiable function $f : \mathbb{R}^n \mapsto [0, \infty)$ is said to be σ-hypoconvex,$\sigma \in \mathbb{R}$ if the following holds:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\sigma}{2} \|y - x\|^2. \tag{2.7}$$

If $\sigma = 0$, (2.7) reduces to the definition of convex function; $\sigma > 0$ corresponds to *strong* convexity whereas when $\sigma < 0$, it is known as the *weak* convexity [35], [43], [46].

A well-known example is the negative entropy function, which is 1-strongly convex in 1-norm [65], and due to the norm relation $\|x\|_2 \leq \|x\|_1, x \in \mathbb{R}^{|\mathcal{X}|}$, it is also 1-strongly convex in 2-norm. Another example is the conditional entropy, which is weakly convex if the corresponding conditional probability mass is ε-infimal as shown in the follow lemma.

Lemma 2.4.3. *Let $G(p_{z|y}) = H(Z|Y)$. If $p_{z|y}$ is a $\epsilon_{z|y}$-infimal measure. Then the function G is $(2N_zN_y/\epsilon_{z|y})$-weakly convex. Where $N_z = |\mathcal{Z}|, N_y = |\mathcal{Y}|$ is the cardinality of the random variables Z, Y respectively.*

Proof. For two arbitrary $p^m_{z|y}, p^n_{z|y} \in \Omega_g$ where the superscript here means two arbitrary iteration count $m, n \in \mathbb{N}$. Then, consider the following:

$$\begin{aligned} H(Z^m|Y) - H(Z^n|Y) &= \sum_y p(y)[\langle p^m_{z|Y} - p^n_{z|Y}, -\log p^m_{z|Y}\rangle - D_{KL}(p^m_{z|Y}||p^n_{z|Y})] \\ &\geq \langle \nabla H(Z^m|Y), p^m_{z|y} - p^n_{z|y}\rangle - E_y\left[\frac{1}{\epsilon_{z|y}}\|p^m_{z|Y} - p^n_{z|Y}\|_1^2\right] \\ &\geq \langle \nabla H(Z^m|Y), p^m_{z|y} - p^n_{z|y}\rangle - \frac{N_{z|y}}{\epsilon_{z|y}}\|p^m_{z|y} - p^n_{z|y}\|_2^2, \end{aligned}$$

where the first inequality follows the reverse Pinsker's inequality [69] which holds when $p_{z|y}$ is $\epsilon_{z|y}$-infimal. And the second inequality is due to norm bound $\|x\|_1 \leq \sqrt{N}\|x\|_2, \forall x \in \mathbb{R}^N$. Then by the definition of weakly convex function we complete the proof. □

A closely related concept to hypo-convexity that we called *restricted weakly convexity* is defined as follows:

Definition 2.4.4. *A function $f : \mathbb{R}^d \mapsto [0, \infty)$, is ω-restricted weakly convex, $\omega > 0$ w.r.t. a matrix $A \in \mathbb{R}^{k\times d}$ if $f \in C^1$ and the following holds:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle - \frac{\omega}{2}\|Ay - Ax\|^2. \tag{2.8}$$

The restricted-weakly convex property is adopted in our earlier work [34] to prove the convergence of an ADMM solver for IB. We further extend the application of restricted weak

convexity to prove the locally linear rate of convergence for non-convex splitting methods. This is based on the data-processing inequality [62], [65] and hence we have the following result:

Lemma 2.4.4. *Assume $p_{z|x}$ is $\epsilon_{z|x}$-infimal. Let $G(p_{z|x}) := -\gamma H(Z|X) + H(Z|Y)$ and $Y - X - Z$ forms a Markov chain. If $0 < \gamma < 1$, then for two $p^m_{z|x}, p^n_{z|x} \in \Omega_{z|x}$, where $\Omega_{z|x} := \{p(z|x) | \sum_z p(z|x) = 1, \forall z \in \mathcal{Z}, x \in \mathcal{X}\}$, $G(p_{z|x})$ is ω_G-restricted weakly convex w.r.t. Q_x.*

$$G(p^m_{z|x}) - G(p^n_{z|x}) \geq \langle \nabla G(p^n_{z|x}), p^m_{z|x} - p^n_{z|x} \rangle - \frac{\omega_G}{2} \|Q_x p^m_{z|x} - Q_x p^n_{z|x}\|^2,$$

where $\omega_G := (2N_z N_x \zeta)/\epsilon_{z|x} - \gamma, \zeta := \sum_y \zeta^2(y)/p(y)$ and $\zeta(y) := \sup_{x \in \mathcal{X}} p(y|x) - \inf_{x \in \mathcal{X}} p(y|x)$.

Proof. As $G(p_{z|x})$ consists of two conditional entropy functions, the proof consists of two parts. For the first part:

$$\begin{aligned}
-H(Z^m|X) + H(Z^n|X) =& \sum_x p(x) \left\{ \sum_z \left[p(z^m|x) - p(z^n|x) \right] \left(\log p(z^n|x) + 1 \right) \right\} \\
&+ E_x[D_{KL}(p^m_{z|X} || p^n_{z|X})] \\
\geq& \langle p^m_{z|x} - p^n_{z|x}, p(x)(\log p^n_{z|x} + 1) \rangle + D_{KL}(Q_x p^m_{z|x} || Q_x p^n_{z|x}) \\
\geq& \langle p^m_{z|x} - p^n_{z|x}, p(x)(\log p^n_{z|x} + 1) \rangle + \|Q_x p^m_{z|x} - Q_x p^n_{z|x}\|_1^2 \\
\geq& \langle p^m_{z|x} - p^n_{z|x}, p(x)(\log p^n_{z|x} + 1) \rangle + \|Q_x p^m_{z|x} - Q_x p^n_{z|x}\|_2^2,
\end{aligned} \tag{2.9}$$

where we use the log-sum inequality for the first and Pinsker's inequality for the second [65] follow by 2-norm bounds. For the second part, without loss of generality, let $\gamma = 1$:

$$\begin{aligned}
H(Z^m|Y) - H(Z^n|Y) =& \sum_y p(y) \left[\sum_z \left(p(z^m|y) - p(z^n|y) \right) \left(-\log p(z^n|y) \right) \right] \\
&- E_y[D_{KL}(p^m_{z|Y} || p^n_{z|Y})] \\
=& \sum_{x,y,z} p(x,y) \left[p(z^m|x) - p(z^n|x) \right] \left[-\log p(z^n|y) \right] - E_y[D_{KL}(p^m_{z|Y} || p^n_{z|Y})] \\
=& \langle p^m_{z|x} - p^n_{z|x}, \nabla_{z|x} H(Z^n|Y) \rangle - E_y[D_{KL}(p^m_{z|Y} || p^n_{z|Y})],
\end{aligned} \tag{2.10}$$

Where $\nabla_{z|x}$ denotes the gradient w.r.t. $p_{z|x}$. Then for the second term of the last line of (2.10), due to $\varepsilon_{z|y}$-infimality we can apply the reverse Pinkser's inequality [69]:

$$E_y\left[D_{KL}[p^m_{z|Y}||p^n_{z|Y}]\right] \leq E_y\left[\frac{N_z}{\varepsilon_{z|y}}\|p^m_{z|Y} - p^n_{z|Y}\|_2^2\right].$$

Then through similar techniques adopted in differential privacy [70], or equivalently, the data-processing inequality with $Q_{x|y}$ as the transition kernel [62], we have:

$$p(z^m|y) - p(z^n|y) \leq \left(\sup_{x\in\mathcal{X}}\frac{p(y|x)}{p(y)} - \inf_{x\in\mathcal{X}}\frac{p(y|x)}{p(y)}\right)\left|\sum_x p(z^m|x)p(x) - p(z^n|x)p(x)\right|. \tag{2.11}$$

Define $\zeta(y) := \sup_{x\in\mathcal{X}} p(y|x) - \inf_{x\in\mathcal{X}} p(y|x)$, substitute (2.11) into (2.10), we have:

$$H(Z^m|Y) - H(Z^n|Y) \geq \langle p^m_{z|x} - p^n_{z|x}, \nabla H(Z^n|Y)\rangle - \frac{N_z N_x}{\epsilon_{z|x}}\left[\sum_y \frac{\zeta^2(y)}{p(y)}\right]\|Q_x p^m_{z|x} - Q_x p^n_{z|x}\|_2^2.$$

Combining the above with γ pre-multiplied to the second part, and then it is clear that $G(p_{z|x})$ satisfies the definite of ω-restricted weakly convex function with $\omega := N_z N_x \zeta/\epsilon_{z|x} - \gamma$ where $\zeta := \sum_y \zeta^2(y)/p(y)$. □

For a smooth function, if it is also convex, then we have the following descent lemma which is commonly used in first-order optimization methods [40], [71], [72].

Definition 2.4.5 (Theorem 2.1.12 [72]). *If $f : \mathbb{R}^n \mapsto [0, +\infty)$ is σ-strongly convex and L-smooth, for some $\sigma, L > 0$, then for any x, y, the following holds:*

$$\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{\sigma L}{\sigma + L}\|x - y\|^2 + \frac{1}{\sigma + L}\|\nabla f(x) - \nabla f(y)\|^2. \tag{2.12}$$

Notably, a recent result generalized Definition 2.4.5 to σ-*hypoconvex* functions f, which can be found in the reference therein [43]. For convenience, in the following, we will assume that p_z is ϵ_z-infimal and $p_{z|x}$ is $\epsilon_{z|x}$-infimal.

With ε-infimal measures, we can follow the standard two-block non-convex ADMM to prove the convergence of the proposed algorithms by showing the corresponding augmented Lagrangian satisfies the KŁ properties. [36], [37], [40]. Once having KŁ properties, the rate

of convergence can be determined in terms of the Łojasiewicz exponent of the augmented Lagrangian.

Definition 2.4.6. *A function $f(x) : R^{|\mathcal{X}|} \mapsto R$ is said to satisfy the Łojasweicz inequality if there exists an exponent $\theta \in [0, 1)$, $\delta > 0$ and a critical point $x^* \in \Omega^*$ with a constant $C > 0$, and a neighborhood $\|x - x^*\| \leq \varepsilon$ such that:*

$$|f(x) - f(x^*)|^\theta \leq C\, dist(0, \nabla f(x)).$$

In literature, there is a broad class of functions known to satisfy the KŁ properties, for example, the o-minimal structure (e.g., sub-analytic, semi-algebraic) [36], [37], [40], [48]. However, in general, verifying a given function satisfies KŁ properties and finding the corresponding Łojasiewicz exponent are difficult tasks [49]. The KŁ inequality characterized the convergence rates around the neighborhood of a stationary point, hence applies to both convex and non-convex functions, in our case, the sacrifice for not having convexity is that the results from KŁ inequality can only assure convergence and the associated rate to local minima.

Definition 2.4.7. *A function $f(x) : R^{|\mathcal{X}|} \mapsto R$ is said to satisfy the Kurdyka-Łojasiewicz inequality if there exists a neighborhood around $\bar{q}$ and a level set $Q := \{q | q \in \Omega, f(q) < f(\bar{q}) < f(q) + \eta\}$ with a margin $\eta > 0$ and a continuous concave function $\varphi(s) : [0, \eta) \to \mathbb{R}_+$, such that the following inequality holds:*

$$\dot{\varphi}(f(q) - f(\bar{q}))\, dist(0, \partial f(q)) \geq 1, \tag{2.13}$$

where ∂f denotes the sub-gradient of $f(\cdot)$ for non-smooth functions, $\dot{\varphi}$ denotes the first-order derivative of φ, and $dist(y, A) := \inf_{x \in \mathcal{A}} \|x - y\|$ is defined as the distance of a set A to a fixed point y if exists.

The following elementary identities are often used in the convergence proof presented in the next section. We list them for self-containing purposes.

$$2\langle u - v, w - u \rangle = \|w - v\|^2 - \|u - v\|^2 - \|u - w\|^2. \tag{2.14}$$

$$\|(1-\alpha)u+\alpha v\|^2 = (1-\alpha)\|u\|^2+\alpha\|v\|^2-\alpha(1-\alpha)\|u-v\|^2. \tag{2.15}$$

Lastly, by "linear" rate of convergence, we refer to the definition in [21]. We note that by the following definition, the Q-linear rate can be considered as a stronger sense of rate of convergence than R-linear, as its decay of the exponent of error is monotonic while it is not in the R-linear case.

Definition 2.4.8. *Let $\{w^k\}$ be a sequence in $\mathbb{R}^n$ that converges to a stationary point w^* when $k > K_0 \in \mathbb{N}$. If it converges Q-linearly, then $\exists Q \in (0,1)$ such that*

$$\frac{\|w^{k+1}-w^*\|}{\|w^k-w^*\|} \leq Q, \quad \forall k > K_0.$$

On the other hand, the convergence of the sequence is R-linear if there is Q-linearly convergent sequence $\{\mu^k\}, \forall k \in \mathbb{N}, \mu^k \geq 0$ such that:

$$\|w^k - w^*\| \leq \mu^k, \forall k \in \mathbb{N}.$$

2.4.2 Proof of Convergence

In proving the convergence of the two algorithms, we consider three different sets of assumptions. We start with the most restricted one since it requires strong convexity of the sub-objective function F and restricted weakly convex on G. However, by exploiting these extra properties, it turns out that they not only have simpler expression in theoretic results but also easier to implement and optimize in practice. Continuing on the proof, the following set of assumptions is paired with the algorithm (2.4):

Assumption A.

- *There exists stationary points $w^* := (p^*, Bq^*, \nu^*)$ that belongs to a set $\Omega^* := \{w | w \in \Omega, \nabla \mathcal{L}_c = 0\}$.*
- *$F(p)$ is L_p-smooth, σ_F-strongly convex while $G(q)$ is L_q-smooth and ω_G-restricted weakly convex.*

- *A is positive definite.*
- *The penalty coefficient $c > c_{\min}$, where $c_{\min}$ is defined as:*

$$c_{\min} := \max\{\omega_G, (L_F + \sigma_F)/(\alpha\mu_A^2)\}.$$

We consider first-order optimization methods, which gives the following minimizer conditions:

$$\begin{aligned}
\nu_{1/2}^{k+1} &= \nu^k - (1-\alpha)\, c\left(Ap^k - Bq^k\right), \\
\nabla F(p^{k+1}) &= -A^T\nu_{1/2}^{k+1} - cA^T\left(Ap^{k+1} - Bq^k\right) \\
&= -A^T\nu^{k+1}, \\
\nu^{k+1} &= \nu_{1/2}^{k+1} + c\left(Ap^{k+1} - Bq^k\right), \\
\nabla G(q^{k+1}) &= B^T[\nu^{k+1} + c\left(Ap^{k+1} - Bq^{k+1}\right)].
\end{aligned} \tag{2.16}$$

Note that at a stationary point $w^* := (p^*, q^*, \nu^*)$ the above reduces to:

$$Ap^* = Bq^*, \quad \nabla F(p^*) = -A^T\nu^*, \quad \nu_{1/2}^* = \nu^*, \quad \nabla G(q^*) = B^T\nu^*. \tag{2.17}$$

With the minimizer conditions shown above, we derive a sufficient decrease lemma for the first algorithm-assumption pair, which gives the next result.

Lemma 2.4.5. *Let $\mathcal{L}_c$ defined as in (2.3) and Assumption A is satisfied, then with algorithm (2.4), we have:*

$$\mathcal{L}_c(p^k, q^k, \nu^k) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu^{k+1}) \geq \delta_p\|p^k - p^{k+1}\|^2 + \delta_q\|Bq^k - Bq^{k+1}\|^2 + \delta_\nu\|\nu^k - \nu^{k+1}\|^2,$$

where the coefficients $\delta_p, \delta_q, \delta_\nu$ are defined as:

$$\delta_p := \frac{\sigma_F L_p}{L_p + \sigma_F} + c\mu_A^2\left(\frac{1}{\alpha} - \frac{1}{2}\right), \quad \delta_q := \frac{c - \omega_G}{2}, \quad \delta_\nu := \frac{\mu_A^2}{L_p + \sigma_F} - \frac{1}{c\alpha},$$

where μ_A denotes the largest and smallest eigenvalue of the positive definite matrix A.

Proof. The proof of the lemma simply follows the four relations below. We start with the relaxation step.

$$\mathcal{L}_c(p^k, q^k, \nu^k) - \mathcal{L}_c(p^k, q^k, \nu_{1/2}^{k+1}) = -(\alpha - 1)\, c\|Ap^k - Bq^k\|^2. \tag{2.18}$$

Then for p-update, due to the σ_F-strong convexity and using Definition 2.4.5, we have:

$$\begin{aligned}
&\mathcal{L}_c(p^k, q^k, \nu_{1/2}^{k+1}) - \mathcal{L}_c(p^{k+1}, q^k, \nu_{1/2}^{k+1}) \\
=&F(p^k) - F(p^{k+1}) + \langle \nu_{1/2}^{k+1}, Ap^k - Ap^{k+1}\rangle + \frac{c}{2}\|Ap^k - Bq^k\|^2 - \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 \\
\geq&\langle \nabla F(p^{k+1}) + A^T\nu_{1/2}^{k+1}, p^k - p^{k+1}\rangle + \frac{c}{2}\|Ap^k - Bq^k\|^2 - \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 \\
&\qquad + \frac{1}{L_p + \sigma_F}\|\nabla F(p^k) - \nabla F(p^{k+1})\|^2 + \frac{\sigma_F L_p}{L_p + \sigma_F}\|p^k - p^{k+1}\|^2 \\
\geq& - c\langle Ap^{k+1} - Bq^k, Ap^k - Ap^{k+1}\rangle + \frac{c}{2}\|Ap^k - Bq^k\|^2 - \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 \\
&\qquad + \frac{\mu_A^2}{L_p + \sigma_F}\|\nu^k - \nu^{k+1}\| + \frac{\sigma_F L_p}{L_p + \sigma_F}\|p^k - p^{k+1}\|^2 \\
=&\frac{c}{2}\|Ap^k - Ap^{k+1}\| + \frac{\sigma_F L_p}{L_p + \sigma_F}\|p^k - p^{k+1}\|^2 + \frac{\mu_A^2}{L_p + \sigma_F}\|\nu^k - \nu^{k+1}\|,
\end{aligned} \tag{2.19}$$

where the first inequality is due to σ_F-strong convexity; the second is due to A is positive definite. Then, for the dual update, we have:

$$\mathcal{L}_c(p^{k+1}, q^k, \nu_{1/2}^{k+1}) - \mathcal{L}_c(p^{k+1}, q^k, \nu^{k+1}) = -c\|Ap^{k+1} - Bq^k\|^2. \tag{2.20}$$

Combine (2.18) and (2.20) using the identity (2.15), we get:

$$\begin{aligned}
&- c\,(\alpha - 1)\,\|Ap^k - Bq^k\|^2 - c\|Ap^{k+1} - Bq^k\|^2 \\
&\qquad = -\frac{1}{c\alpha}\|\nu^{k+1} - \nu^k\|^2 - c\left(1 - \frac{1}{\alpha}\right)\|Ap^k - Ap^{k+1}\|^2.
\end{aligned} \tag{2.21}$$

Lastly, for the q-update, since G is ω_G-restricted weakly convex w.r.t. the matrix B:

$$\begin{aligned}
&\mathcal{L}_c(p^{k+1}, q^k, \nu^{k+1}) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu^{k+1}) \\
=&G(q^k) - G(q^{k+1}) + \langle \nu^{k+1}, Bq^{k+1} - Bq^k \rangle + \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 - \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2. \\
\geq&\langle \nabla G(q^{k+1}) - B^T\nu^{k+1}, q^k - q^{k+1} \rangle - \frac{\omega_G}{2}\|Bq^k - Bq^{k+1}\|^2 + \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 \\
&\qquad - \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 \\
=&c\langle Ap^{k+1} - Bq^{k+1}, Bq^k - Bq^{k+1} \rangle - \frac{\omega_G}{2}\|Bq^k - Bq^{k+1}\|^2 + \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 \\
&\qquad - \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 \\
=&\frac{c - \omega_G}{2}\|Bq^k - Bq^{k+1}\|^2,
\end{aligned} \tag{2.22}$$

Summing (2.19)(2.21)(2.22), and using (2.16), we have:

$$\begin{aligned}
&\mathcal{L}_c(p^k, q^k, \nu^k) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu^{k+1}) \\
&\qquad \geq \left[\frac{\mu_A^2}{L_p + \sigma_F} - \frac{1}{c\alpha}\right]\|\nu^k - \nu^{k+1}\|^2 + \frac{c - \omega_G}{2}\|Bq^k - Bq^{k+1}\|^2 \\
&\qquad\qquad + c\left(\frac{1}{\alpha} - \frac{1}{2}\right)\|Ap^k - Ap^{k+1}\|^2 + \frac{\sigma_F L_p}{L_p + \sigma_F}\|p^k - p^{k+1}\|^2,
\end{aligned} \tag{2.23}$$

Then by positive definiteness of A we have: $\|Ap^k - Ap^{k+1}\| \geq \mu_A\|p^k - p^{k+1}\|$, where μ_A denotes the smallest eigenvalue of A. Substitute this into the above and we complete the proof.

□

By Lemma 2.4.5, the conditions that assure sufficient decrease are equivalent to the range of the penalty coefficient c and the relaxation parameter α such that the coefficients $\delta_p, \delta_q, \delta_\nu$ are non-negative. When the conditions are satisfied, the sufficient decrease lemma implies the convergence of the algorithm (2.4).

Lemma 2.4.6. *Suppose Assumption A is satisfied and $0 < \alpha \leq 2$. Define the collective point at step k as $w^k := (p^k, Bq^k, \nu^k)$, the sequence $\{w^k\}_{k\in\mathbb{N}}$ obtained from the algorithm (2.4) is convergent to a stationary point $w^* \in \Omega^*$.*

Proof. By Assumption A, the coefficients $\delta_p, \delta_q, \delta_\nu$ defined in Lemma 2.4.6 are non-negative, so the next step is to show $\{\mathcal{L}_c^k\}_{k\in\mathbb{N}}$ is finite. Denote $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$ for simplicity. From the above, assume a penalty coefficient c^* satisfies Assumption A, we have:

$$\sum_{k=1}^{N-1} \mathcal{L}_c^k - \mathcal{L}_c^{k+1} = \mathcal{L}_c^1 - \mathcal{L}_c^N \geq C^* \sum_{k=1}^{N} \left[\|p^k - p^{k+1}\|^2 + \|\nu^k - \nu^{k+1}\|^2 + \|Bq^k - Bq^{k+1}\|^2 \right], \tag{2.24}$$

where $C^* = \min\{\delta_p, \delta_q, \delta_\nu\} > 0$. Define the collective point at step k as $w^k := (p^k, Bq^k, \nu^k)$, then since there exist stationary points w^*, the l.h.s. of (2.24) is lower semi-continuous. Let $N \to \infty$ and denote the limit point w^∞, since $\mathcal{L}_c^1 - \mathcal{L}_c^\infty$ is finite, the r.h.s. of (2.24) is finite. This implies $\|w^k - w^{k+1}\|^2 \to 0$ as $k \to \infty$, since $\sum^\infty \|w^k - w^{k+1}\|^2$ is a Cauchy sequence. From this, we know that $w^\infty \in \Omega^*$, or equivalently, for $k > N_0 \in \mathbb{N}$ sufficiently large, $w^k \to w^*$ as $k \to \infty$, which proves that $\{w^k\}_{k\in\mathbb{N}}$ is convergent to w^*.

□

As a remark, the convergence is not point-wise. This can be observed as q in the collective point $w^k = (p^k, Bq^k, \nu^k)$ is pre-multiplied by the matrix B. In practice, take IB for example, this corresponds to the symmetry of solutions [73], [74]. Nonetheless, in the information-theoretic optimization considered, the point-wise convergence is not necessary as the mutual information is invariant to symmetry.

As mentioned earlier, observe that Lemma 2.4.6 requires the function F to be strongly convex and G be restricted weakly convex with respect to the matrix B. It turns out these requirements can be relaxed. To see this, consider the second algorithm (2.5) paired with the assumptions shown below. We can similarly develop a sufficient decrease lemma for this alternative algorithm-assumption pair.

Assumption B.

- *There exists stationary points* $w^* := (Ap^*, q^*, \nu^*)$ *that belong to a set* $\Omega^* := \{w | w \in \Omega, \nabla\mathcal{L}_c = 0\}$,
- *The function* $F(p)$ *is* L_p*-smooth, convex while* $G(q)$ *is* L_q*-smooth,* σ_G*-weakly convex.*

- *B is positive definite; A is full row rank.*
- *The penalty coefficient c satisfies:*

$$c > \frac{\alpha\sigma_G + \sqrt{\alpha^2\sigma_G^2 + 8(2-\alpha)L_q^2}}{(4-2\alpha)\mu_B^2}.$$

With the corresponding first-order minimizer conditions:

$$\begin{aligned}
\nabla F(p^{k+1}) &= -A^T[\nu^k + c(Ap^{k+1} - Bq^k)], \\
\nu_{1/2}^{k+1} &= \nu^k - (1-\alpha)\, c\left(Ap^{k+1} - Bq^k\right), \\
\nabla G(q^{k+1}) &= B^T[\nu_{1/2}^{k+1} + c\left(Ap^{k+1} - Bq^{k+1}\right)] \\
&= B^T\nu^{k+1}, \\
\nu^{k+1} &= \nu_{1/2}^{k+1} + c\left(Ap^{k+1} - Bq^{k+1}\right).
\end{aligned} \tag{2.25}$$

Lemma 2.4.7. *Let $\mathcal{L}_c$ defined as in (2.3) and Assumption B is satisfied, then with the algorithm (2.5), we have:*

$$\begin{aligned}
\mathcal{L}_c(p^k, q^k, \nu^k) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu^{k+1}) \geq \frac{c}{2}\|Ap^k - Ap^{k+1}\|^2 - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 \\
+ c\left(\frac{1}{\alpha} - \frac{1}{2}\right)\|Bq^k - Bq^{k+1}\|^2 - \frac{1}{\alpha c}\|\nu^k - \nu^{k+1}\|^2.
\end{aligned}$$

Proof. First, by assumption F is convex and hence:

$$\begin{aligned}
&\mathcal{L}_c(p^k, q^k, \nu^k) - \mathcal{L}_c(p^{k+1}, q^k, \nu^k) \\
=&F(p^k) - F(p^{k+1}) + \langle \nu^k, Ap^k - Ap^{k+1}\rangle + \frac{c}{2}\|Ap^k - Bq^k\|^2 - \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 \\
\geq&\langle \nabla F(p^{k+1}) + A^T\nu^k, p^k - p^{k+1}\rangle + \frac{c}{2}\|Ap^k - Bq^k\|^2 - \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 \\
=&-c\langle Ap^{k+1} - Bq^k, Ap^k - Ap^{k+1}\rangle + \frac{c}{2}\|Ap^k - Bq^k\|^2 - \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 \\
=&\frac{c}{2}\|Ap^k - Ap^{k+1}\|^2,
\end{aligned} \tag{2.26}$$

where the last equality is due to the minimizer conditions (2.25). Then for the relaxation step (2.5b):

$$\mathcal{L}_c(p^{k+1}, q^k, \nu^k) - \mathcal{L}_c(p^{k+1}, q^k, \nu_{1/2}^{k+1}) = -(\alpha - 1)\, c\|Ap^{k+1} - Bq^k\|^2. \tag{2.27}$$

On the other hand, by assumption, G is σ_G-weakly convex, so we have the following lower bound for q-update (2.5c):

$$\begin{aligned}
&\mathcal{L}_c(p^{k+1}, q^k, \nu_{1/2}^k) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu_{1/2}^k) \\
=&G(q^k) - G(q^{k+1}) - \langle \nu_{1/2}^{k+1}, Bq^k - Bq^{k+1}\rangle + \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 - \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 \\
\geq&\langle \nabla G(q^{k+1}) - B^T\nu_{1/2}^{k+1}, q^k - q^{k+1}\rangle - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 + \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 \\
&- \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 \\
=&c\langle Ap^{k+1} - Bq^{k+1}, Bq^k - Bq^{k+1}\rangle - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 + \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2 \\
&- \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 \\
=&\frac{c}{2}\|Bq^k - Bq^{k+1}\|^2 - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2.
\end{aligned} \tag{2.28}$$

Lastly, for the dual ascend (2.5d):

$$\mathcal{L}_c(p^{k+1}, q^{k+1}, \nu_{1/2}^{k+1}) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu^{k+1}) = -c\|Ap^{k+1} - Bq^{k+1}\|^2. \tag{2.29}$$

Combine the above using the identity (2.15), we get:

$$\begin{aligned}
\frac{1}{c\alpha}\|\nu^k - \nu^{k+1}\|^2 &= \|c\left(Ap^{k+1} - Bq^{k+1}\right) - c\,(1-\alpha)\,[Ap^{k+1} - Bq^k]\|^2 \\
&= c\|Ap^{k+1} - Bq^{k+1}\|^2 + c\,(\alpha - 1)\,\|Ap^{k+1} - Bq^k\|^2 - c\left(1 - \frac{1}{\alpha}\right)\|Bq^k - Bq^{k+1}\|^2,
\end{aligned} \tag{2.30}$$

Summing (2.26)(2.28)(2.30), we get:

$$\mathcal{L}_c(p^k, q^k, \nu^k) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu^{k+1}) \geq \frac{c}{2}\|Ap^k - Ap^{k+1}\|^2 - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 + c\left(\frac{1}{\alpha} - \frac{1}{2}\right)\|Bq^k - Bq^{k+1}\|^2 - \frac{1}{\alpha c}\|\nu^k - \nu^{k+1}\|^2, \tag{2.31}$$

which completes the proof. □

Similar to the convergence results Lemma 2.4.6 implied by the sufficient decrease lemma (Lemma 2.4.5), we have the following convergence result for the second algorithm (2.5) based on Lemma 2.4.7.

Lemma 2.4.8. *Suppose Assumption B is satisfied and* $0 < \alpha < 2$. *Define* $w^k := (Ap^k, q^k, \nu^k)$ *the collective point at step k. Then the sequence* $\{w^k\}_{k\in\mathbb{N}}$ *obtained from the algorithm* (2.5) *is convergent to a stationary point* $w^* \in \Omega^*$.

Proof. By assumption, B is positive definite, denote its smallest eigenvalue μ_B and $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$ for simplicity, we have:

$$\|q^k - q^{k+1}\| = \|(B^{-1}B)q^k - q^{k+1}\| \leq \|B^{-1}\|\|Bq^k - Bq^{k+1}\|,$$

Note that $\|B^{-1}\| = 1/\mu_B$. On the other hand, for the dual variable, we have:

$$\|\nu^k - \nu^{k+1}\| = \|(B^{-T}B^T)(\nu^k - \nu^{k+1})\| \leq \|B^{-T}\|\|\nabla G(q^k) - \nabla G(q^k)\| \leq L_q\|B^{-T}\|\|q^k - q^{k+1}\|,$$

combining the two results above, we have the following lower bound to Lemma 2.4.7:

$$\mathcal{L}_c^k - \mathcal{L}_c^{k+1} \geq \left[c\mu_B^2(\frac{1}{\alpha} - \frac{1}{2}) - \frac{\sigma_G}{2} - \frac{L_q^2}{\alpha c \mu_B^2}\right]\|q^k - q^{k+1}\|^2 + \frac{c}{2}\|Ap^k - Ap^{k+1}\|^2.$$

To make the scalar pre-multiplied $\|q^k - q^{k+1}\|^2$ be positive, it follows from elementary quadratic programming, which gives the range of the penalty coefficient c to be:

$$c > \frac{\alpha\sigma_G + \sqrt{\alpha^2\sigma_G^2 + 8(2-\alpha)L_q^2}}{(4-2\alpha)\mu_B^2},$$

which is satisfied as listed in Assumption B. Denote $c^* > 0$ that satisfies the above condition, we have:

$$\mathcal{L}_c^k - \mathcal{L}_c^{k+1} \geq c^* \left[\|q^k - q^{k+1}\|^2 + \|Ap^k - Ap^{k+1}\|^2 \right].$$

Then, consider the following:

$$\sum_{i=1}^{N-1} \mathcal{L}_c^k - \mathcal{L}_c^{k+1} = \mathcal{L}_c^1 - \mathcal{L}_c^N \geq c^* \sum_{i=1}^{N-1} \left[\|q^k - q^{k+1}\|^2 + \|Ap^k - Ap^{k+1}\|^2 \right]. \tag{2.32}$$

Define the collective point at step k as $w^k := (Ap^k, q^k, \nu^k)$, since there exist stationary points $w^* := (Ap^*, q^*, \nu^*)$ by assumption, the l.h.s. of (2.32) is lower semi-continuous. Note that the the r.h.s of (2.32) does not depends on the dual variable ν, we can further define a condensed point at step k as $z^k := (Ap^k, q^k)$. Then, by letting $N \to \infty$ and denote the limit point z^∞, since $\mathcal{L}_c^1 - \mathcal{L}_c^\infty$ is finite, the r.h.s. of (2.32) is finite. This implies $\|z^k - z^{k+1}\|^2 \to 0$ as $k \to \infty$, since $\sum^\infty \|z^k - z^{k+1}\|^2$ is a Cauchy sequence. From this we know that $z^\infty = z^*$. Moreover, by (2.25):

$$\sqrt{\mu_{BB^T}} \|\nu^k - \nu^{k+1}\| \leq \|B^T\nu^k - B^T\nu^{k+1}\| = \|\nabla G(q^k) - \nabla G(q^{k+1})\| \leq L_q \|q^k - q^{k+1}\|. \tag{2.33}$$

Hence $L_q^2 \mu_{BB^T} \|q^k - q^{k+1}\|^2 \geq \|\nu^k - \nu^{k+1}\|^2$ and hence $\|\nu^k - \nu^{k+1}\|^2 \to 0$ as $k \to \infty$. In turns, we know $\nu^\infty = \nu^*$. So, together we have $w^k \to w^*$ as $k \to \infty$ which proves that $\{w^k\}_{k\in\mathbb{N}}$ is convergent to w^*. □

Note that the convergence of the second algorithm (2.5) requires no strong-convexity for the sub-objective function $F(p)$, Moreover, the assumption on $G(q)$ is more relaxed than that of the algorithm 2.4, hence potentially applies to more general problems. However, in practice, the more relaxed conditions result in more variables to optimize.

Another major difference between Assumption A and B lies in the linear constraints. In Assumption A, A is positive definite while B is positive definite in Assumption B instead. In the Markovian information theoretic optimization problem, we considered (2.1), the linear constraints $Ap - Bq$ are interpreted as the marginal/Markov relations of (conditional) probabilities. Therefore, only one of the two matrices A, B is an identity matrix, while the

other will be singular. Then for problems such as PF, whose convex sub-objective function is not strongly convex, with A being positive definite instead of B, neither Assumption A nor Assumption B applies. Inspired by [45], where G is Lipschitz continuous and smooth, we can relax Assumption A but keep the A to be positive definite as in Assumption B. This addresses the previously mentioned technical difficulty, but we hence need the third set of assumptions, which is summarized below.

Assumption C.

- *There exists stationary points* $w^* := (p^*, q^*, \nu^*)$ *that belongs to a set* $\Omega^* := \{w | w \in \Omega, \nabla \mathcal{L}_c = 0\}$,
- *The function* $F(p)$ *is* L_p*-smooth, convex while* $G(q)$ *is* L_q*-smooth,* σ_G*-weakly convex,*
- *In addition,* $G(q)$ *is* M_q*-Lipschitz continuous,*
- A *is positive definite;* B *is full row rank,*
- *The penalty coefficient* c *satisfies:*

$$c > M_q \left[\frac{M_q \alpha \sigma_G + \sqrt{M_q^2 \alpha^2 \sigma_G^2 + 8(2-\alpha) L_q^2 \lambda_B^2 / \mu_{BB^T}}}{4 - 2\alpha} \right].$$

When the above assumptions are imposed on the algorithm (2.5), which reuses the minimization conditions (2.25), we have the following sufficient decrease lemma.

Lemma 2.4.9. *Suppose Assumption C is satisfied and* $0 < \alpha < 2$. *Define* $w^k := (p^k, q^k, \nu^k)$ *the collective point at step* k, *then the sequence* $\{w^k\}_{k \in \mathbb{N}}$ *obtained from the algorithm* (2.5) *is convergent to a stationary point* $w^* \in \Omega^*$.

Proof. Following the steps (2.26)(2.28)(2.30), we start from (2.31).

Define $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$ the function value evaluated with variables at step k for simplicity:

$$\begin{aligned}\mathcal{L}_c^k - \mathcal{L}_c^{k+1} \geq & \frac{c}{2}\|Ap^k - Ap^{k+1}\|^2 - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 + c\left(\frac{1}{\alpha} - \frac{1}{2}\right)\|Bq^k - Bq^{k+1}\|^2 \\ & - \frac{1}{\alpha c}\|\nu^k - \nu^{k+1}\|^2 \\ \geq & \frac{c\mu_A^2}{2}\|p^k - p^{k+1}\|^2 + \left[\frac{c}{M_q^2}\left(\frac{1}{\alpha} - \frac{1}{2}\right) - \left(\frac{\sigma_G}{2} + \frac{L_q^2\lambda_B^2}{\alpha c \mu_{BB^T}}\right)\right]\|q^k - q^{k+1}\|^2,\end{aligned} \tag{2.34}$$

where in the last inequality, the first term is by A being positive definite, and for the second term, we follow [45] and use Lipschitz continuity of G to have $\|q^k - q^{k+1}\| \leq M_q\|Bq^k - Bq^{k+1}\|$; we denote $\lambda_B := \|B\|$ as the largest positive singular value of a matrix B and μ_B for the smallest positive eigenvalue of B; For $\|\nu^k - \nu^{k+1}\|$, since B is full row rank and G is L_q-smooth, we have:

$$\begin{aligned}\|\nu^k - \nu^{k+1}\| &= \|(BB^T)^{-1}BB^T(\nu^k - \nu^{k+1})\| \\ &\leq \frac{\lambda_B}{\sqrt{\mu_{BB^T}}}\|\nabla G(q^k) - \nabla G(q^{k+1})\| \\ &\leq \frac{\lambda_B}{\sqrt{\mu_{BB^T}}}L_q\|q^k - q^{k+1}\|.\end{aligned} \tag{2.35}$$

From elementary quadratic programming, the range in terms of the penalty coefficient c that assures the second term of the last inequality in (2.34) is positive:

$$c > M_q\left[\frac{M_q\sigma_G\alpha + \sqrt{(M_q\sigma_G\alpha)^2 + 8(2-\alpha)L_q^2\lambda_B^2/\mu_{BB^T}}}{4 - 2\alpha}\right]$$

Then by assumption, c satisfies the above. Rewrite the coefficients as $\tau_p, \tau_q > 0$ for simplicity, then there exists a $\tau^* := \min\{\tau_p, \tau_q\}$ such that:

$$\mathcal{L}_c^k - \mathcal{L}_c^{k+1} \geq \tau^*\left(\|p^k - p^{k+1}\|^2 + \|q^k - q^{k+1}\|^2\right),$$

Then denote $w^k := (p^k, q^k, \nu^k)$ the collective point at step k; $\mathcal{L}_c^k := \mathcal{L}_c(w^k)$ the function value evaluated with w^k. Summing both sides of the inequality (2.35), we have:

$$\sum_{k=1}^{N-1} \mathcal{L}_c^k - \mathcal{L}_c^{k+1} = \mathcal{L}_c^1 - \mathcal{L}_c^N \geq \tau^* \sum_{k=1}^{N-1} \left(\|p^k - p^{k+1}\|^2 + \|q^k - q^{k+1}\|^2 \right),$$

By assumption the l.h.s. of the above inequality is lower semi-continuous and therefore is finite. So as $N \to \infty$, $\mathcal{L}_c^1 - \mathcal{L}_c^\infty < +\infty$. This implies the r.h.s. is finite and therefore $\|p^k - p^{k+1}\|^2 \to 0$ and $\|q^k - q^{k+1}\|^2 \to 0$ as $k \to \infty$. Due to (2.35), we know that $\|\nu^k - \nu^{k+1}\|^2 \to 0$ as well. Given the results, denote the limit points as $w^\infty := (p^\infty, q^\infty, \nu^\infty)$, since $\|w^k - w^{k+1}\|^2 \to 0$ as $k \to \infty$, $w^\infty = w^*$ which proves that $\{w^k\}_{k \in \mathbb{N}}$ is convergent to w^*. □

2.4.3 Rate of Convergence Analysis

In this part, based on the convergence results derived in the last section, we further prove that the rate of convergence of three algorithm-assumption pairs are all achieving locally linear. It turns out that the linear convergence we derived is independent of initialization and the sequence obtained from the associated algorithm converges to local minimizers when the current update of the variables lies around the neighborhood of them. This is consistent with the recent results [45], [75], based on the KŁ inequality that characterize the rate of convergence for splitting methods in a broad class of non-convex problems [36], [37], [40], [71]. Overall, to adopt the KŁ inequality, it consists of two steps. First, we explicitly show that the associated Łojasiewicz exponents $\theta = 1/2$ of (2.3), solved with the two algorithms (2.4) and (2.5). And the two cases both satisfy the KŁ properties. Then we apply the following result, owing to [36], [40], [49], and hence prove the linear convergence rate.

Lemma 2.4.10 (Theorem 2 [36]). *Assume that a function $\mathcal{L}_c(p, q, \nu)$ satisfies the KŁ properties, define w^k the collective point at step k, and let $\{w^k\}_{k \in \mathbb{N}}$ be a sequence generated by either the alternating algorithm (2.4) or (2.5). Suppose $\{w^k\}_{k \in \mathbb{N}}$ is bounded and the following relation holds:*

$$\|\nabla \mathcal{L}_c^k\| \leq C^* \|w^k - w^{k-1}\|,$$

where $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$ *and* $C^* > 0$ *is some constant. Denote the Łojasiewicz exponent of* $\mathcal{L}_c$ *with* $\{w^\infty\}$ *as* θ*. Then the following holds:*

(i) If $\theta = 0$*, the sequences* $\{w^k\}_{k\in\mathbb{N}}$ *converges in a finite number of steps,*

(ii) If $\theta \in (0, 1/2]$ *then there exist* $\tau > 0$ *and* $Q \in [0, 1)$ *such that*

$$|w^k - w^\infty| \leq \tau Q^k,$$

(iii) If $\theta \in (1/2, 1)$ *then there exists* $\tau > 0$ *such that*

$$|w^k - w^\infty| \leq \tau k^{-\frac{1-\theta}{2\theta-1}},$$

Proof. We only prove the case corresponds to $\theta = 1/2$ as it is relevant to the following discussion. The proof for other scenarios is referred to [36].

$\mathcal{L}_c(p, q, \nu)$ satisfies the KŁ properties with an exponent $\theta = 1/2$.

Denote $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$, without loss of generality let $\mathcal{L}_c^* = 0$ and define a concave function $\Phi(s) := C_0 s^{1-\theta}$ with $C_0, s > 0$. For $k > N_0 \in \mathbb{N}$ sufficiently large, by the concavity of Φ (Note that the gradient is evaluated with w^k):

$$\begin{aligned}(\mathcal{L}_c^k)^{1-\theta} - (\mathcal{L}_c^{k+1})^{1-\theta} &\geq (1-\theta)\,(\mathcal{L}_c^k)^{-\theta}\left[\mathcal{L}_c^k - \mathcal{L}_c^{k+1}\right] \\ &\geq C\,(1-\theta)\,(\mathcal{L}_c^k)^{-\theta}\|w^{k+1} - w^k\|^2 \\ &\geq C\,(1-\theta)\,\|\nabla\mathcal{L}_c^k\|^{-1}\|w^{k+1} - w^k\|^2,\end{aligned} \tag{2.36}$$

where $C > 0$ due to Lemma 2.4.6; and the last inequality is due to Lemma 2.4.14.

Then, by assumption, for some constant $C^* > 0$, we have:

$$\|\nabla\mathcal{L}_c^k\| \leq C^*\|w^k - w^{k-1}\|. \tag{2.37}$$

Substitute the above into (2.36), define $C_1 := C/C^*(1-\theta)$, we get:

$$(\mathcal{L}_c^k)^{1-\theta} - (\mathcal{L}_c^{k+1})^{1-\theta} \geq C_1\frac{\|w^{k+1} - w^k\|^2}{\|w^k - w^{k-1}\|}.$$

Substitute the above into (2.37), we have:

$$2\|w^{k+1} - w^k\| \leq \|w^k - w^{k-1}\| + C_2 \left[(\mathcal{L}_c^k)^{1-\theta} - (\mathcal{L}_c^{k+1})^{1-\theta} \right] \tag{2.38}$$

where we define $C_2 := \sqrt{1/(2C_1)}$. For the first inequality, we use the identity $2ab \leq a^2 + b^2$; the second inequality is due to the non-increasing sequence $\{\mathcal{L}_c^k\}_{k\in\mathbb{N}}$; the third inequality is due to the KŁ properties, and the last inequality follows (2.37). Then, by defining $\Delta_k := \sum_{l=k}^{\infty} \|w^{l+1} - w^l\|$, and summing both sides of (2.38) with $k \in \mathbb{N}$, we have:

$$\Delta_k \leq (\Delta_{k-1} - \Delta_k) + C_4 \left(\Delta_{k-1} - \Delta_k\right)^{\frac{1-\theta}{\theta}}. \tag{2.39}$$

Finally, from Lemma 2.4.14, $\theta = 1/2$, we have $(1-\theta)/\theta = 1$ and therefore:

$$\Delta_k \leq \frac{K^*}{1+K^*} \Delta_{k-1},$$

where $K^* = 1 + C_4 > 0$. The above proves the locally linear rate of convergence. Specifically, Q-linearly fast in terms of the Cauchy sequence Δ_k. □

The above result characterizes the rate of convergence in terms of the KŁ exponent, but except for certain types of functions, the calculation of the KŁ exponent is difficult. The following key result, due to [49], is useful in calculating the KŁ exponent of (2.3) and is included for completeness.

Lemma 2.4.11 (Lemma 2.1 [49]). *Suppose that f is a proper closed function, $\nabla f(\bar{w}) \neq 0$. Then for any $\theta \in [0, 1)$, f satisfies the KŁ properties at $\bar{w}$ with an exponent of θ there exists $\eta := \frac{1}{2}\|\nabla f(\bar{w})\| > 0$ and $\delta \in (0, 1)$ such that $\|\nabla f(w)\| > \eta$ whenever $\|w - \bar{w}\| \leq \varepsilon$ and $f(\bar{w}) < f(w) < f(\bar{w}) + \delta$.*

In literature, the KŁ inequality has been successfully adopted to find the rate of convergence for alternating algorithms such as ADMM and recently PRS or DRS with $\alpha = (1+\sqrt{5})/2$. For more general DRS methods in terms of the relaxation parameter α, we find that proving locally linear rate through the KŁ inequality only holds for $1 \leq \alpha \leq 2$. As

for $0 < \alpha < 1$, inspired by the recent results that shows locally R-linear rate of convergence for the primal ADMM [47], we adopt and extend the approach to the two algorithms (2.4) and (2.5) under the three sets of assumptions. Combining the two methods, we therefore theoretically prove that the rates are locally linear for $0 < \alpha \leq 2$.

Lemma 2.4.12. *Let $\mathcal{L}_c$ defined as in (2.3) and let the sequence $\{w^k\}_{k\in\mathbb{N}}$ obtained through either the algorithm (2.4) or (2.5) is bounded. Denote $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$. Suppose the following holds for some $K^* > 0$:*

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^k \leq K^* \left[\mathcal{L}_c^k - \mathcal{L}_c^{k+1} + \|w^{k+1} - w^*\|^2\right],$$

and there exists a neighborhood around a stationary point w^*, such that $\|w - w^*\| < \epsilon$, $\mathcal{L}_c^* < \mathcal{L}_c < \mathcal{L}_c^* + \delta$ with $\delta, \epsilon > 0$. Then $\{\mathcal{L}_c^k\}_{k\in\mathbb{N}}$ is Q-linearly convergent and $\{w^k\}_{k\in\mathbb{N}}$ converges R-linearly to w^* around the neighborhood.

Proof. By assumption, denote $\Delta_c^k := \mathcal{L}_c^k - \mathcal{L}_c^*$, we have:

$$\Delta_c^{k+1} \leq K^* \left[(\Delta_c^k - \Delta_c^{k+1}) + \|w^{k+1} - w^*\|^2\right].$$

Then around a neighborhood of w^*, we get:

$$\frac{\Delta_c^{k+1}}{\Delta_c^k} < \frac{K^*}{1+K^*} + \frac{K^*\epsilon^2}{1+K^*}\left(\frac{1}{\Delta_c^k}\right) \leq \frac{K^*}{1+K^*} + \frac{K^*\epsilon^2}{1+K^*}\left(\frac{1}{\Delta_c^{k+1}}\right) < \frac{K*}{1+K^*} + \frac{K^*\epsilon^2/\xi}{1+K^*},$$

where the second inequality follows from the sufficient descent lemma and by definition, $\delta > \mathcal{L}_c^{k+1} - \mathcal{L}_c^* > \xi > 0$, as $w^{k+1} \notin \Omega^*$. Therefore, we can simply choose $\epsilon < \sqrt{\xi/K^*} < \sqrt{\delta/K^*}$, which shows the convergence of the sequence of function values $\{\mathcal{L}_c^k\}_{k\in\mathbb{N}}$ is Q-linear locally around the neighborhood of a stationary point w^*. In turns, we have for $n > N_0 \in \mathbb{N}$:

$$\begin{aligned}
\rho_p\|Ap^n - Ap^{n+1}\|^2 \leq \mathcal{L}_c^n - \mathcal{L}_c^{n+1} \leq K_p Q^n,\\
\rho_q\|q^n - q^{n+1}\|^2 \leq \mathcal{L}_c^n - \mathcal{L}_c^{n+1} \leq K_q Q^n,\\
\rho_\nu\|\nu^n - \nu^{n+1}\|^2 \leq \mathcal{L}_c^n - \mathcal{L}_c^{n+1} \leq K_\nu Q^n,
\end{aligned}$$

for some $K_p, K_q, K_\nu > 0$ and $0 < Q < 1$. Combine the above together, we have:

$$\bar{\rho}\|w^n - w^{n+1}\|^2 \leq \bar{K}Q^n,$$

where $\bar{K} = K_p + K_q + K_\nu$ and $\bar{\rho} = \min\{\rho_p, \rho_q, \rho_\nu\}$. Now, for the sequence $\{w^n\}_{n\in\mathbb{N}\setminus[N_0]}$ around w^*, by taking $m > n \geq N_0$, we have:

$$\|w^n - w^m\|^2 \leq \sum_{\mathrm{i}=n}^{m}\|w^n - w^{n+1}\|^2 \leq \frac{\bar{K}Q^n}{\bar{\rho}(1-Q)}.$$

Since the above is a Cauchy sequence, by taking limit with $m \to \infty$, which gives $w^m \to w^*$ as $m \to \infty$, we get:

$$\|w^n - w^*\|^2 \leq \frac{\bar{K}Q^n}{\bar{\rho}(1-Q)},$$

and hence proving that $\{w^n\}_{n>N_0}$ is R-linearly convergent. □

Remarkably, the rate of convergence with KŁ inequality is Q-linear, or in other words, monotonic convergence in terms of the error between variables $\|w^k - w^{k-1}\|$ in consecutive steps while the R-linear rate is non-monotonic, hence a weaker rate. However, the weaker R-linear rate comes with milder assumptions imposed on the linear constraints as the full row rank assumptions are lifted.

In the rest of this part, we aim at proving the sequence $\{w^k\}_{k\in\mathbb{N}}$ obtained from the proposed two algorithms both satisfy the KŁ properties. The results are based on the following lemmas. We start with that correspond to the first algorithm (2.4).

Lemma 2.4.13. *Let $\mathcal{L}_c$ defined as in (2.3). For the sequence $\{w^k\}_{k\in\mathbb{N}}$ obtained from the algorithm (2.4) where $w^k := (p^k, Bq^k, \nu^k)$, if it is bounded and converges to a stationary point w^* satisfying (2.17), then we have:*

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \left(\frac{c\lambda_A^2}{2} - \frac{\sigma_F L_p}{L_p + \sigma_F}\right)\|p^{k+1} - p^*\|^2 - \frac{c - \omega_G}{2}\|Bq^{k+1} - Bq^*\|^2,$$

and

$$\|\nabla\mathcal{L}_c(w^{k+1})\|^2 \geq \left(c^2\mu_A^2 + 1\right)\|Ap^{k+1} - Bq^{k+1}\|^2,$$

where $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$*;* λ_A, μ_A *denote the largest and smallest eigenvalue of a positive definite matrix* A*.*

Proof. By the definition (2.3), the properties of F and G, following algorithm (2.4) with the first order minimizer conditions (2.16), and denote $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$ for simplicity, we have:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c =& F(p^{k+1}) + G(q^{k+1}) + \langle \nu^{k+1}, Ap^{k+1} - Bq^{k+1}\rangle + \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 \\
&- F(p) - G(q) - \langle \nu, Ap - Bq\rangle - \frac{c}{2}\|Ap - Bq\|^2 \\
\leq& \langle \nabla F(p^{k+1}), p^{k+1} - p\rangle - \frac{\sigma_F L_p}{L_p + \sigma_F}\|p^{k+1} - p\|^2 + \langle \nabla G(q^{k+1}), q^{k+1} - q\rangle \\
&+ \frac{\omega_G}{2}\|Bq^{k+1} - Bq\|^2 + \langle \nu^{k+1}, Ap^{k+1} - Bq^{k+1}\rangle - \langle \nu, Ap - Bq\rangle \\
&+ \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 - \frac{c}{2}\|Ap - Bq\|^2 \\
=& \langle \nu^{k+1} - \nu, Ap - Bq\rangle + c\langle Ap^{k+1} - Bq^{k+1}, Bq^{k+1} - Bq\rangle + \frac{\omega_G}{2}\|Bq^{k+1} - Bq\|^2 \\
&- \frac{\sigma_F L_p}{L_p + \sigma_F}\|p^{k+1} - p\|^2 + \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 - \frac{c}{2}\|Ap - Bq\|^2,
\end{aligned} \tag{2.40}$$

where the first inequality is due to Definition 2.4.5 and restricted-weak convexity of G. Substituting w^* to w, that is $Ap^* = Bq^*$, and using identity (2.14) for the second inner product in the last line of (2.40), we get:

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \left(\frac{c\lambda_A^2}{2} - \frac{\sigma_F L_p}{L_p + \sigma_F}\right)\|p^{k+1} - p^*\|^2 - \frac{c - \omega_G}{2}\|Bq^{k+1} - Bq^*\|^2,$$

where λ_A denotes the largest eigenvalue of the matrix A. Then, for the second part, consider the following:

$$\nabla \mathcal{L}_c^{k+1} = \begin{bmatrix} \nabla F(p^{k+1}) + A^T[\nu^{k+1} + c\left(Ap^{k+1} - Bq^{k+1}\right)] \\ \nabla G(q^{k+1}) - B^T[\nu^{k+1} + c\left(Ap^{k+1} - Bq^{k+1}\right)] \\ Ap^{k+1} - Bq^{k+1} \end{bmatrix} = \begin{bmatrix} cA^T\left(Ap^{k+1} - Bq^{k+1}\right) \\ 0 \\ Ap^{k+1} - Bq^{k+1} \end{bmatrix},$$

where the last equality follows from (2.16). By showing that $\|\nabla \mathcal{L}_c^{k+1}\|^2 \geq K\|Ap^{k+1} - Bq^{k+1}\|^2$ with $K := c^2\mu_A^2 + 1$ where μ_A is the smallest eigenvalue of the matrix A, we complete the proof. □

Lemma 2.4.14. *Suppose Assumption A is satisfied, if the augmented Lagrangian (2.3) is solved with the algorithm (2.4), then it satisfies KŁ properties with an exponent $\theta = 1/2$.*

Proof. Using Lemma 2.4.13, we simply add an additional positive squared norm $\|Ap^{k+1} - Bq^{k+1}\|^2$:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* &\leq \left(\frac{c\lambda_A^2}{2} - \frac{\sigma_F L_p}{L_p + \sigma_F}\right)\|p^{k+1} - p^*\|^2 - \frac{c - \omega_G}{2}\|Bq^{k+1} - Bq^*\|^2 \\
&\leq \left(\frac{c\lambda_A^2}{2} - \frac{\sigma_F L_p}{L_p + \sigma_F}\right)\|p^{k+1} - p^*\|^2 \\
&\leq \left(\frac{c\lambda_A^2}{2} - \frac{\sigma_F L_p}{L_p + \sigma_F}\right)\|p^{k+1} - p^*\|^2 + \|Ap^{k+1} - Bq^{k+1}\|^2 \\
&= c_G\|p^{k+1} - p^*\|^2 + \|Ap^{k+1} - Bq^{k+1}\|^2,
\end{aligned}$$

where we define $c_G := (c\lambda_A^2)/2 - (\sigma_F L_p)/(L_p + \sigma_F)$. Then we have:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq c_G\|p^{k+1} - p^*\|^2 + \|Ap^{k+1} - Bq^{k+1}\|^2 &\leq c_G\epsilon^2 + \frac{1}{K_1}\|\nabla \mathcal{L}_c^{k+1}\|^2 \\
&\leq \|\nabla \mathcal{L}_c^{k+1}\|^2\left[\frac{c_G\epsilon^2}{\eta^2} + \frac{1}{K_1}\right],
\end{aligned}$$

where $K_1 := c^2\mu_A^2 + 1 > 0$; the first inequality is due to Lemma 2.4.13 and $\|w^{k+1} - w^*\| < \epsilon$ around the neighborhood of w^*; the last inequality follows from Lemma 2.4.11. By taking square root of both sides, we complete the proof. □

From Lemma 2.4.14, the Lojasiewicz exponent $\theta = 1/2$. By mapping the exponent according to Lemma 2.4.10, we can show the linear rate of convergence. Furthermore, we can combine with the convergence (Lemma 2.4.6) together to have the following result.

Theorem 2.4.15. *Suppose Assumption A is satisfied. For $0 < \alpha \leq 2$, define $w^k := (p^k, Bq^k, \nu^k)$ the collective point at step k. Then the sequence $\{w^k\}_{k\in\mathbb{N}}$ obtained from the*

algorithm (2.4) *is bounded. Moreover, the sequence converges to a stationary point* w^* *at linear rate locally.*

Proof. The convergence follows the sufficient decrease lemma (Lemma 2.4.6), so it suffices to prove the rate of convergence. By assumption, the penalty coefficient is sufficiently large such that Lemma 2.4.6 holds. In addition to convergence, for the corresponding rate, due to Lemma 2.4.14, $\mathcal{L}_c(p, q, \nu)$ satisfies the KŁ properties with an exponent $\theta = 1/2$. For the gradient norm $\|\nabla \mathcal{L}_c\|$, by Lemma 2.4.13 we have:

$$\|\nabla \mathcal{L}_c^k\| \leq c_a \|Ap^k - Bq^k\| \leq c_a \left(\|Ap^k - Bq^{k-1}\| + \|Bq^k - Bq^{k-1}\| \right), \tag{2.41}$$

where $c_a := c + \lambda_A$. Then, suppose $1 \leq \alpha \leq 2$, by (2.21):

$$\|Ap^k - Bq^{k-1}\|^2 \leq \frac{1}{\alpha^2 c^2} \|\nu^k - \nu^{k-1}\|^2 + \left(1 - \frac{1}{\alpha}\right) \|Ap^k - Ap^{k-1}\|^2.$$

Substitute the above into (2.41), we get:

$$\begin{aligned} \|\nabla \mathcal{L}_c^k\| &\leq c_\alpha \left[\|Bq^k - Bq^{k-1}\| + \left(\frac{1}{\alpha^2 c^2} \|\nu^k - \nu^{k-1}\|^2 + \left(1 - \frac{1}{\alpha}\right) \|Ap^k - Ap^{k-1}\|^2 \right)^{\frac{1}{2}} \right] \\ &\leq c_\alpha^* \left[\|\nu^k - \nu^{k-1}\| + \|p^k - p^{k-1}\| + \|Bq^k - Bq^{k-1}\| \right] \\ &= c_\alpha^* \|w^k - w^{k-1}\|, \end{aligned} \tag{2.42}$$

where $c_\alpha^* := \max\{c_\alpha/(\alpha^2 c^2), c_\alpha \lambda_A\}$. Then, following similar steps in (2.42), we conclude that, for $1 \leq \alpha \leq 2$ and some constant $c_{\alpha t} > 0$, we have:

$$\|\nabla \mathcal{L}_c^k\| \leq c_{\alpha t} \|w^k - w^{k-1}\|.$$

Then by Lemma 2.4.10, we prove the locally linear rate of convergence for the case $1 \leq \alpha \leq 2$. On the other hand, for $0 < \alpha < 1$, from Lemma 2.4.6 and Assumption A, there exists a constant $K^* > 0$ such that:

$$\mathcal{L}_c^k - \mathcal{L}_c^{k+1} \geq K^* \|w^k - w^{k+1}\|^2,$$

Moreover, denote $w^* := (p^*, Bq^*, \nu^*)$ a stationary point. Due to Lemma 2.4.13, we have:

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \left(\frac{c\lambda_A^2}{2} - \frac{\sigma_F L_p}{L_p + \sigma_F} \right) \|p^{k+1} - p^*\|^2 - \frac{c - \omega_G}{2} \|Bq^{k+1} - Bq^k\|^2.$$

By Assumption A, there always exists $K_1 > 0$ and a neighborhood around the stationary point w^* such that:

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq K_1 \|w^{k+1} - w^*\|^2 \leq K_1 \|w^k - w^*\|^2 + \frac{1}{K^*} \|w^{k+1} - w^k\| \leq K_1 \|w^{k+1} - w^*\|^2 + \mathcal{L}_c^k - \mathcal{L}_c^{k+1}.$$

Then by Lemma 2.4.12, we conclude that the sequence $\{w^k\}_{k > \mathbb{N}_0}, N_0 \in \mathbb{N}$ converges R-linearly to w^*. This completes the proof for linear rate of convergence for the full range of $0 < \alpha \leq 2$.

□

Similarly, for the second algorithm (2.5), we can show that the Łojasiewicz exponent of the corresponding augmented Lagrangian is $\theta = 1/2$ and the KŁ properties are satisfied. However, it turns out that applying KŁ for this alternative algorithm-assumption pair requires the matrix A to be full row rank. We later show that this additional assumption is not necessary to prove locally linear in an alternative approach, but a weaker sense R-linear rate as a result.

Lemma 2.4.16. *Let $\mathcal{L}_c$ defined as in (2.3). For the sequence $\{w^k\}_{k \in \mathcal{N}}$ obtained from the algorithm (2.5) where $w^k := (Ap^k, q^k, \nu^k)$, if it is bounded and converges to a stationary point w^* satisfying (2.17), $0 < \alpha < 2$, then we have:*

$$\begin{aligned} \mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq\ & c\|Ap^{k+1} - Bq^{k+1}\|^2 + \frac{\sigma_G}{2}\|q^{k+1} - q^*\|^2 - \frac{c}{2}\|Bq^{k+1} - Bq^*\|^2 \\ & + \frac{c(2-\alpha)}{2}\|Bq^k - Bq^*\|^2 - \frac{c(2-\alpha)}{2}\|Ap^{k+1} - Bq^k\|^2 + \frac{c(\alpha-1)}{2}\|Ap^{k+1} - Ap^*\|^2, \end{aligned}$$

where $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$. Moreover, if $AA^T \succ 0$, then:

$$\|\nabla \mathcal{L}_c(w^{k+1})\|^2 \geq \mu_{AA^T} \left[\|\nu^k - \nu^{k+1}\|^2 + c^2 \|Bq^k - Bq^{k+1}\|^2 - 2cL_q \|q^k - q^{k+1}\|^2 \right],$$

where μ_W denotes the smallest positive eigenvalue of a matrix W.

Proof. For the first part, denote $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$ the function value evaluate with the variables at step k, we have:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c =& F(p^{k+1}) + G(q^{k+1}) + \langle \nu^{k+1}, Ap^{k+1} - Bq^{k+1}\rangle + \frac{\sigma_G}{2}\|q^{k+1} - q\|^2 \\
&+ \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 - F(p) - G(q) - \langle \nu, Ap - Bq\rangle - \frac{c}{2}\|Ap - Bq\|^2 \\
\leq& \langle \nabla F(p^{k+1}), p^{k+1} - p\rangle + \langle \nabla G(q^{k+1}), q^{k+1} - q\rangle + \langle \nu^{k+1}, Ap^{k+1} - Bq^{k+1}\rangle \\
&+ \frac{\sigma_G}{2}\|q^{k+1} - q\|^2 - \langle \nu, Ap - Bq\rangle + \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 - \frac{c}{2}\|Ap - Bq\|^2 \\
=& \langle \nu^{k+1} - \nu, Ap - Bq\rangle + \frac{c}{2}\|Ap^{k+1} - Bq^{k+1}\|^2 - \frac{c}{2}\|Ap - Bq\|^2 + \frac{\sigma_G}{2}\|q^{k+1} - q\|^2 \\
&+ c\langle Ap^{k+1} - Bq^{k+1}, Ap^{k+1} - Ap\rangle - c(2-\alpha)\langle Ap^{k+1} - Bq^k, Ap^{k+1} - Ap\rangle,
\end{aligned} \tag{2.43}$$

where the first inequality follows from convexity of F and weak convexity of G. By assumption $0 < \alpha \leq 2$, we have:

$$\begin{aligned}
& c\langle Ap^{k+1} - Bq^{k+1}, Ap^{k+1} - Ap\rangle - c(2-\alpha)\langle Ap^{k+1} - Bq^k, Ap^{k+1} - Ap\rangle \\
=& \frac{c}{2}\left[-\|Bq^{k+1} - Ap\|^2 + \|Ap^{k+1} - Bq^{k+1}\|^2 + \|Ap^{k+1} - Ap\|^2\right] - \frac{c(2-\alpha)}{2}\left[-\|Ap - Bq^k\|^2\right. \\
&\left. + \|Ap^{k+1} - Bq^k\|^2 + \|Ap^{k+1} - Ap\|^2\right].
\end{aligned} \tag{2.44}$$

Substitute (2.44) into (2.43), using identities (2.14)(2.15), and let $w := w^*$, which gives $Ap^* = Bq^*$, we have:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq& c\|Ap^{k+1} - Bq^{k+1}\|^2 - \frac{c}{2}\|Bq^{k+1} - Bq^*\|^2 + \frac{\sigma_G}{2}\|q^{k+1} - q^*\|^2 \\
&+ \frac{c(\alpha-1)}{2}\|Ap^{k+1} - Ap^*\|^2 + \frac{c(2-\alpha)}{2}\|Bq^k - Bq^*\|^2 - \frac{c(2-\alpha)}{2}\|Ap^{k+1} - Bq^k\|^2.
\end{aligned} \tag{2.45}$$

By identity (2.15) and the minimizer conditions (2.25):

$$c\|Ap^{k+1} - Bq^{k+1}\|^2 = \frac{1}{c\alpha}\|\nu^{k+1} - \nu^k\|^2 - c(\alpha-1)\|Ap^{k+1} - Bq^k\|^2 + c\left(1 - \frac{1}{\alpha}\right)\|Bq^k - Bq^{k+1}\|^2.$$

Substitute the above into (2.45), we have:

$$\begin{aligned}\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq & \frac{1}{c\alpha}\|\nu^k - \nu^{k+1}\|^2 + \frac{\sigma_G}{2}\|q^{k+1} - q^*\|^2 - \frac{c}{2}\|Bq^{k+1} - Bq^*\|^2 \\ & + \frac{c(\alpha-1)}{2}\|Ap^{k+1} - Ap^*\|^2 + \frac{c(2-\alpha)}{2}\|Bq^k - Bq^*\|^2 \\ & - \frac{c\alpha}{2}\|Ap^{k+1} - Bq^k\|^2 + c\left(1 - \frac{1}{\alpha}\right)\|Bq^k - Bq^{k+1}\|^2,\end{aligned} \tag{2.46}$$

and we complete the proof for the first part. For the second part, consider:

$$\begin{aligned}\nabla\mathcal{L}_c^{k+1} &= \begin{bmatrix} \nabla F(p^{k+1}) + A^T\left[\nu^{k+1} + c(Ap^{k+1} - Bq^{k+1})\right] \\ \nabla G(q^{k+1}) - B^T\left[\nu^{k+1} + c(Ap^{k+1} - Bq^{k+1})\right] \\ Ap^{k+1} - Bq^{k+1} \end{bmatrix} \\ &= \begin{bmatrix} A^T[\nu^{k+1} - \nu^k + c(Bq^k - Bq^{k+1})] \\ -cB^T(Ap^{k+1} - Bq^{k+1}) \\ Ap^{k+1} - Bq^{k+1} \end{bmatrix}.\end{aligned} \tag{2.47}$$

Denote the smallest positive eigenvalue of a matrix W as μ_W, by assumption, since $AA^T \succ 0$, we have:

$$\begin{aligned}\|\nabla\mathcal{L}_c^{k+1}\|^2 \geq & \mu_{AA^T}\left[\|\nu^k - \nu^{k+1}\|^2 + c^2\|Bq^k - Bq^{k+1}\|^2 - 2c\langle\nu^k - \nu^{k+1}, Bq^k - Bq^{k+1}\rangle\right] \\ & + \left(c^2\mu_{BB^T} + 1\right)\|Ap^{k+1} - Bq^{k+1}\|^2 \\ \geq & \mu_{AA^T}\left[\|\nu^k - \nu^{k+1}\|^2 + c^2\|Bq^k - Bq^{k+1}\|^2 - 2cL_q\|q^k - q^{k+1}\|^2\right].\end{aligned} \tag{2.48}$$

where in the last inequality, we use the minimizer condition (2.25) and L_q-smoothness of G. □

Lemma 2.4.17. *Suppose Assumption B is satisfied and the matrix A is full row rank. For $0 < \alpha < 2$, if the augmented Lagrangian (2.3) is solved with the algorithm (2.5), then it satisfies KŁ inequality with an exponent $\theta = 1/2$.*

Proof. From Lemma 2.4.16, bounding the terms with negative coefficients from the above with 0 and using Cauchy-Schwarz inequality on $\|Bq^k - Bq^*\|^2$, denote $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$, we have:

$$\begin{aligned}\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq & \frac{1}{\alpha c}\|\nu^k - \nu^{k+1}\|^2 + c\left(3 - \alpha - \frac{1}{\alpha}\right)\|Bq^k - Bq^{k+1}\|^2 \\ & + \frac{c(\alpha-1)}{2}\|Ap^{k+1} - Ap^*\|^2 + \left[\frac{\sigma_G}{2} - c\lambda_B^2\left(\frac{1}{2\mu_B^2} + 2 - \alpha\right)\right]^+ \|q^{k+1} - q^*\|^2,\end{aligned} \tag{2.49}$$

where the first inequality follows applying Cauchy-Schwarz inequality $\|u+v\|^2 \leq 2(\|u\|^2 + \|v\|^2)$ to $\|Bq^k - Bq^*\|^2$. Note that for the coefficient of the term $\|Bq^k - Bq^{k+1}\|^2$, it follows $\alpha + 1/\alpha \geq 2$. Then by defining $2C_G := [\sigma_G - c\lambda_B^2(1/\mu_B^2 + 4 - 2\alpha)]^+$ where $[\,\cdot\,]^+ := \max\{0, \cdot\}$, we have:

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \frac{1}{\alpha c}\|\nu^k - \nu^{k+1}\|^2 + c\lambda_B^2\|q^k - q^{k+1}\|^2 + \frac{c|\alpha-1|}{2}\|Ap^{k+1} - Ap^*\|^2 + C_G\|q^{k+1} - q^*\|. \tag{2.50}$$

On the other hand, since B is positive definite by assumption, we can further find a lower bound of (2.48):

$$\begin{aligned}\|\nabla\mathcal{L}_c^{k+1}\|^2 \geq \mu_{AA^T}\left[\|\nu^k - \nu^{k+1}\|^2 + \left(c^2\mu_B^2 - 2cL_q\right)\|q^k - q^{k+1}\|^2\right] \\ \geq K_1\left(\|\nu^k - \nu^{k+1}\|^2 + \|q^k - q^{k+1}\|^2\right).\end{aligned}$$

If we further assume $c > 2L_q/\mu_B^2$ and define $K_1 := \mu_{AA^T}\min\{1, c^2 - 2cL_q\}$. Combining the above, then there always exists a scalar $K_2 := \max\{1/(\alpha c), c\lambda_B^2\}$ such that:

$$\begin{aligned}\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq & K_2\left(\|\nu^k - \nu^{k+1}\|^2 + \|q^k - q^{k+1}\|^2\right) + \frac{c|\alpha-1|}{2}\|Ap^{k+1} - Ap^*\|^2 + C_G\|q^{k+1} - q^*\|^2 \\ \leq & \frac{K_2}{K_1}\|\nabla\mathcal{L}_c^{k+1}\|^2 + K_3\left(\|Ap^{k+1} - Ap^*\|^2 + \|q^{k+1} - q^*\|^2 + \|\nu^{k+1} - \nu^*\|^2\right) \\ = & \frac{K_2}{K_1}\|\nabla L_c^{k+1}\|^2\left(1 + \frac{K_3\|w^{k+1} - w^*\|^2}{\|\nabla\mathcal{L}_c^{k+1}\|^2}\right) \\ \leq & \frac{K_2}{K_1}\|\nabla L_c^{k+1}\|^2\left(1 + \frac{K_3\varepsilon^2}{\eta^2}\right),\end{aligned}$$

where $K_3 := \max\{c|\alpha - 1|/2, C_G\}$ and the last inequality follows Lemma 2.4.11, that is, around a neighborhood of w^* with $\|w - w^*\| < \varepsilon$ there exists a $\eta > 0$ such that $\|\nabla \mathcal{L}_c^{k+1}\| > \eta$. By taking square root of both sides of the above, we conclude that the Łojasiewicz exponent $\theta = 1/2$, which completes the proof. □

As mentioned previously, in Lemma 2.4.17, the matrix A is required to be a full row rank matrix. This is necessary to prove Q-linear rate of convergence with KŁ inequality. It turns out that we can relax this condition by showing local linear rate of convergence without assuming A being full row rank.

To see this, observe that Lemma 2.4.18 implies that the sequence $\{\mathcal{L}_c^k\}_{k\in\mathbb{N}}$ is locally Q-linear convergent, which in turns allow us to show locally R-linear rate of convergence of the sequence $\{w^k\}_{k\in\mathbb{N}}$ obtained from the second algorithm (2.5).

Theorem 2.4.18. *Suppose Assumption B is satisfied and the sequence $\{w^k\}_{k\in\mathbb{N}}$ with $w^k := (Ap^k, q^k, \nu^k)$ obtained from the algorithm 2.5 is bounded, then the sequence $\{w^k\}$ converges linearly toward a stationary point w^* locally around its neighborhood $\|w - w^*\|^2 < \epsilon$ and $\mathcal{L}_c^* < \mathcal{L}_c < \mathcal{L}_c^* + \eta$ for some $\epsilon, \eta > 0$.*

Proof. Denote $\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)$ for simplicity. From Lemma 2.4.7, there always exists a stationary point $w^* := (Ap^*, q^*, \nu^*)$ where the sequence $\{\mathcal{L}_c^k\}$ is converging to. By assumption, the penalty coefficient c is large enough such the sufficient decrease lemma holds, which proves the convergence. For the corresponding rate, for $1 \leq \alpha < 2$, by Lemma 2.4.17, the KŁ exponent $\theta = 1/2$. In addition, from (2.47) we have the following:

$$\begin{aligned}\|\nabla \mathcal{L}_c^{k+1}\|^2 &\leq \|A^T[\nu^{k+1} - \nu^k + c(Bq^k - Bq^{k+1})]\|^2 + (1 + c^2\lambda_B^2)\|Ap^{k+1} - Bq^{k+1}\|^2 \\ &\leq 2\lambda_{AA^T}(\|\nu^{k+1} - \nu^k\|^2 + c^2\|Bq^{k+1} - Bq^k\|^2) + (1 + c\lambda_B^2)\|Ap^{k+1} - Bq^{k+1}\|^2,\end{aligned}$$

where the inequality follows that A is full row rank by assumption. Then, using the identity (2.15) and minimizer conditions (2.25), we have:

$$\|Ap^{k+1} - Bq^{k+1}\|^2 + (\alpha - 1)\|Ap^{k+1} - Bq^k\|^2 = \frac{1}{c^2\alpha}\|\nu^{k+1} - \nu^k\|^2 + \left(1 - \frac{1}{\alpha}\right)\|Bq^{k+1} - Bq^k\|^2.$$

Since $1 \leq \alpha < 2$, we have the following:

$$\|\nabla \mathcal{L}_c^{k+1}\|^2 \leq \left(2\lambda_{AA^T} + \frac{1+c\lambda_B^2}{c^2\alpha}\right)\|\nu^{k+1}-\nu^k\|^2 + \left[2c^2\lambda_{AA^T} + \left(1+c\lambda_B^2\right)\left(1-\frac{1}{\alpha}\right)\right]\|Bq^{k+1}-Bq^k\|^2.$$

For the second term, since B is assumed to be positive definite, $\|Bq^{k+1}-Bq^k\|^2 \leq \lambda_B^2\|q^{k+1}-q^k\|^2$. Substitute into the above and define $M^* := \max\{2\lambda_{AA^T}+(1+c\lambda_B^2)/(c^2\alpha), \lambda_B^2[2c^2\lambda_{AA^T}+(1+c\lambda_B^2)(1+1/\alpha)]\}$, we have:

$$\|\nabla \mathcal{L}_c^{k+1}\|^2 \leq M^*\|w^{k+1}-w^k\|^2. \tag{2.51}$$

Substitute (2.37) with (2.51), then by Lemma 2.4.10, we prove that the sequence $\{w^k\}_{k>N_0}$ for some $N_0 \in \mathbb{N}$ converges Q-linearly to w^* around its neighborhood. On the other hand, for $0 < \alpha < 1$, from (2.50), we have:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* &\leq \frac{1}{\alpha c}\|\nu^k-\nu^{k+1}\|^2 + c\lambda_B^2\|q^k-q^{k+1}\|^2 + \frac{c|\alpha-1|}{2}\|Ap^{k+1}-Ap^*\|^2 \\
&\qquad + C_G\|q^{k+1}-q^*\| \\
&\leq \left[\frac{L_q^2}{\mu_B^2\alpha c} + c\lambda_B^2\right]\|q^k-q^{k+1}\|^2 + \frac{c|\alpha-1|}{2}\|Ap^{k+1}-Ap^*\|^2 + C_G\|q^{k+1}-q^*\|^2 \\
&\leq \left[\frac{L_q^2}{\mu_B^2\alpha c} + c\lambda_B^2\right]\|q^k-q^{k+1}\|^2 \\
&\qquad + C^*\left(\|Ap^{k+1}-Ap^*\|^2 + \|q^{k+1}-q^*\|^2 + \|\nu^{k+1}-\nu^*\|^2\right),
\end{aligned}$$

where $2C_G := [\sigma_G - c\lambda_B^2(1/\mu_B^2 + 4 - 2\alpha)]^+$, $C^* := \max\{C_G, c|\alpha-1|/2\}$ and the second inequality is due to the L_q-smoothness of the sub-objective function G. On the other hand, from Lemma 2.4.7, we have:

$$\mathcal{L}_c^k - \mathcal{L}_c^* - (\mathcal{L}_c^{k+1} - \mathcal{L}_c^*) \geq \left[\frac{c}{\mu_B^2}\left(\frac{1}{\alpha}-\frac{1}{2}\right) - \frac{\sigma_G}{2} - \frac{\mu_B^2 L_q^2}{c\alpha}\right]\|q^k-q^{k+1}\|^2 = K_G\|q^k-q^{k+1}\|^2,$$

where by assumption $K_G > 0$. Then by combining the above two results and defining the constants $C_q := \mu_B^2 L_q^2/(\alpha c) + c\lambda_B^2/2$, $K^* := \max\{C_q/K_G, C^*\}$, denote $\Delta^k := \mathcal{L}_c^k - \mathcal{L}_c^*$, we have:

$$\Delta^{k+1} \leq K^* \left(\Delta^k - \Delta^{k+1}\right) + C^* \|w^{k+1} - w^*\|^2 \leq \max\{K^*, C^*\} \left[\left(\Delta^k - \Delta^{k+1}\right) + \|w^{k+1} - w^*\|^2\right].$$

Then by Lemma 2.4.12, the sequence $\{w^k\}_{k>N_0}, N_0 \in \mathbb{N}$ converges R-linearly to w^*. Therefore, combining the results for $1 \leq \alpha < 2$ and $0 < \alpha < 1$ together, we conclude that the rate of convergence for $0 < \alpha < 2$ is locally linear. □

Lastly, for the algorithm (2.5), under the Assumption C, we follow the same framework. First, we prove the Łojasiewicz exponent in solving the augmented Lagrangian (2.3) is $\theta = 1/2$. Second, we apply the KŁ inequality to show its convergence rate is Q-linear.

Compared to previous cases, the main difference of the Assumption C is that the sub-objective function G is further required to be Lipschitz continuous. If this additional condition holds, then the inequality such as $\|q^m - q^n\| \leq M_q \|Bq^m - Bq^n\|$ with B not necessary be positive definite can be shown [45]. In the next result, we also first adopt this technique to show $\theta = 1/2$.

Lemma 2.4.19. *Suppose Assumption C is satisfied and the matrix B is full row rank. For $0 < \alpha < 2$, if the augmented Lagrangian (2.3) is solved with the algorithm (2.5), then it satisfies KŁ inequality with an exponent $\theta = 1/2$.*

Proof. By construction, $\mathcal{L}_c$ is solved with the algorithm (2.5), from (2.46) of Lemma 2.4.16, we have:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq & \frac{1}{c\alpha}\|\nu^k - \nu^{k+1}\|^2 + \frac{\sigma_G}{2}\|q^{k+1} - q^*\|^2 - \frac{c}{2}\|Bq^{k+1} - Bq^*\|^2 \\
& + \frac{c(\alpha-1)}{2}\|Ap^{k+1} - Ap^*\|^2 + \frac{c(2-\alpha)}{2}\|Bq^k - Bq^*\|^2 - \frac{c\alpha}{2}\|Ap^{k+1} - Bq^k\|^2 \\
& + c(1-\frac{1}{\alpha})\|Bq^k - Bq^{k+1}\|^2, \\
\leq & \frac{1}{\alpha c}\|\nu^k - \nu^{k+1}\|^2 + c\lambda_B^2\left(3 - \frac{1}{\alpha} - \alpha\right)\|q^{k+1} - q^k\|^2 + \frac{\sigma_G}{2}\|q^{k+1} - q^*\|^2 \\
& + c\left(\frac{3}{2} - \alpha\right)\|Bq^{k+1} - Bq^*\|^2 + \frac{c\lambda_A^2|\alpha-1|}{2}\|p^{k+1} - p^*\|^2 \\
\leq & \frac{1}{\alpha c}\|\nu^k - \nu^{k+1}\|^2 + c\lambda_B^2\|q^{k+1} - q^k\|^2 + \left[\frac{\sigma_G}{2} + c\lambda_B^2\left|\frac{3}{2} - \alpha\right|\right]\|q^{k+1} - q^*\|^2 \\
& + \frac{c\lambda_A^2|\alpha-1|}{2}\|p^{k+1} - p^*\|^2,
\end{aligned}$$

where the second inequality follows from applying the Cauchy-Schwarz inequality $\|u+v\|^2 \leq 2(\|u\|^2 + \|v\|^2)$ on the term $\|Bq^k - Bq^*\|^2$; the last inequality is because $\alpha + 1/\alpha \geq 2$. Then, by defining $W_1 := \max\{1/(\alpha c), c\lambda_B^2\}$ and $W_G := \max\{\sigma_G/2 + c\lambda_B^2|3/2 - \alpha|, c\lambda_A^2|\alpha - 1|/2\}$, we have:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* &\leq W_1(\|\nu^k - \nu^{k+1}\|^2 + \|q^{k+1} - q^k\|^2) + W_G(\|q^{k+1} - q^*\|^2 + \|p^{k+1} - p^*\|^2) \\
&\leq W_1(\|\nu^k - \nu^{k+1}\|^2 + \|q^{k+1} - q^k\|^2) + W_G\|w^{k+1} - w^*\|^2,
\end{aligned} \tag{2.52}$$

where $w^k := (p^k, q^k, \nu^k)$ denotes the collective point at step k. On the other hand, for the lower bound of $\|\nabla\mathcal{L}_c^{k+1}\|$, from (2.48) we have:

$$\begin{aligned}
\|\nabla\mathcal{L}_c^{k+1}\|^2 \geq & \mu_A^2\left[\|\nu^k - \nu^{k+1}\|^2 + c^2\|Bq^k - Bq^{k+1}\|^2 - 2cL_q\|q^k - q^{k+1}\|^2\right] \\
\geq & \mu_A^2\left[\|\nu^k - \nu^{k+1}\|^2 + (\frac{c^2}{M_q^2} - 2cL_q)\|q^k - q^{k+1}\|^2\right],
\end{aligned}$$

where the second inequality follows Lipschitz continuity of G and (2.5c), which gives $\|q^m - q^n\| \leq M_q\|Bq^m - Bq^n\|, \forall m, n \in \mathbb{N}$ as shown in [45]. Then if we further assume $c > 2L_q/\mu_{BB^T}$ and define $W_2 := \mu_A^2 \min\{1, c^2/M_q^2 - 2cL_q\}$, we get:

$$\|\nabla \mathcal{L}_c^{k+1}\|^2 \geq W_2 \left(\|\nu^k - \nu^{k+1}\|^2 + \|q^k - q^{k+1}\|^2\right). \tag{2.53}$$

Combining (2.52) and (2.53), we have:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* &\leq \frac{W_1}{W_2}\|\nabla \mathcal{L}_c^{k+1}\|^2 + W_G\|w^{k+1} - w^*\|^2 \\
&\leq W^*\|\nabla \mathcal{L}_c^{k+1}\|^2(1 + \frac{\|w^{k+1} - w^*\|^2}{\|\nabla \mathcal{L}_c^{k+1}\|^2}) \\
&\leq W^*\|\nabla \mathcal{L}_c^{k+1}\|^2(1 + \frac{\varepsilon^2}{\eta^2}),
\end{aligned}$$

where $W^* := \max\{W_1/W_2, W_G\}$; the last inequality is due to Lemma 2.4.11 and the definite of a local neighborhood of a stationary point w^*. Finally, by taking square root on both sides of the above inequality, we prove that the Łojasiewicz exponent $\theta = 1/2$. □

Theorem 2.4.20. *Suppose Assumption C is satisfied. For $0 < \alpha < 2$, define $w^k := (p^k, q^k, \nu^k)$ the collective point at step k. Then the sequence $\{w^k\}_{k\in\mathbb{N}}$ obtained from the algorithm (2.5) is bounded. Moreover, the sequence converges to a stationary point w^* at linear rate locally.*

Proof. The convergence of the sequence $\{w^k\}_{k\in\mathbb{N}}$ is due to the sufficient decrease lemma (Lemma 2.4.9) and Assumption C. Moreover, by Lemma 2.4.19, $\mathcal{L}_c$ satisfies the KŁ properties with an exponent $\theta = 1/2$. As in (2.36), we have for some constant $C > 0$:

$$(\mathcal{L}_c^k)^{1-\theta} - (\mathcal{L}_c^{k+1})^{1-\theta} \geq C(1-\theta)\|\nabla \mathcal{L}_c^k\|^{-1}\|w^{k+1} - w^k\|^2.$$

Then from (2.47), we have:

$$\|\nabla \mathcal{L}_c^k\|^2 \leq \lambda_A^2\left[\|\nu^k - \nu^{k-1}\|^2 + (2cL_q + c^2\lambda_B^2)\|q^k - q^{k-1}\|^2\right] + (c^2\lambda_{BB^T} + 1)\|Ap^k - Bq^k\|^2. \tag{2.54}$$

Recall the following, due to the algorithm (2.5) and the identity (2.15):

$$\|Ap^k - Bq^k\|^2 = \frac{1}{c^2\alpha}\|\nu^k - \nu^{k-1}\|^2 + (1 - \frac{1}{\alpha})\|Bq^k - Bq^{k-1}\|^2 - (\alpha - 1)\|Ap^k - Bq^{k-1}\|^2.$$

If $1 \leq \alpha < 2$, substitute the above into (2.54), we have:

$$\begin{aligned}\|\nabla \mathcal{L}_c^k\|^2 \leq & \lambda_A^2 \left[\|\nu^k - \nu^{k-1}\|^2 + (2cL_q + c^2\lambda_B^2)\|q^k - q^{k-1}\|^2\right] \\ & + (c^2\lambda_{BB^T} + 1)\left[\frac{1}{c^2\alpha}\|\nu^k - \nu^{k-1}\|^2 + (1 - \frac{1}{\alpha})\|Bq^k - Bq^{k-1}\|^2\right] \\ \leq & [\lambda_A^2 + \frac{(c^2\lambda_{BB^T} + 1)}{c^2\alpha}]\|\nu^k - \nu^{k-1}\|^2 + [\lambda_A^2 c(2L_q + c\lambda_B^2) + \lambda_B^2(1 - \frac{1}{\alpha})]\|q^k - q^{k-1}\|^2.\end{aligned}$$

Define $S^* := \max\{\lambda_A^2 + (c^2\lambda_{BB^T} + 1)/(c^2\alpha), \lambda_A^2 c(2L_q + c\lambda_B^2) + \lambda_B^2(1 - 1/\alpha)\}$, we have:

$$\|\nabla \mathcal{L}_c^k\| \leq S^*\|w^k - w^{k-1}\|. \tag{2.55}$$

Then, by Lemma 2.4.10 with (2.37) replaced by (2.55), we prove that the rate of convergence for the case $1 \leq \alpha < 2$ is Q-linear. On the other hand, for $0 < \alpha < 1$, by assumption, the following holds, for some constant $\tau^* > 0$, due to Lemma 2.4.9:

$$\mathcal{L}_c^k - \mathcal{L}_c^{k+1} \geq \tau^*(\|p^k - p^{k+1}\|^2 + \|q^k - q^{k+1}\|^2).$$

In addition, from (2.46) with negative terms replaced with 0, we have:

$$\begin{aligned}\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq & \frac{1}{c\alpha}\|\nu^k - \nu^{k+1}\|^2 + \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 + \frac{c(2-\alpha)}{2}\|Bq^k - Bq^*\|^2 \\ \leq & \frac{1}{c\alpha}\|\nu^k - \nu^{k+1}\|^2 + \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 + c(2-\alpha)\|Bq^k - Bq^{k+1}\|^2 \\ & + c(2-\alpha)\|Bq^{k+1} - Bq^*\|^2 \\ \leq & \left[\frac{\mu_{BB^T}\lambda_B L_q^2}{c\alpha} + \frac{\sigma_G}{2} + c\lambda_B^2(2-\alpha)\right]\|q^k - q^{k+1}\|^2 + c\lambda_B^2(2-\alpha)\|q^{k+1} - q^*\|^2.\end{aligned}$$

The second line is by Cauchy-Schwarz inequality, that is, $\|Bq^k - Bq^*\|^2 \leq 2(\|Bq^k - Bq^{k+1}\|^2 + \|Bq^{k+1} - Bq^*\|^2)$, and the third line is by L_q-smooth of G and full row rank assumption of B. Define $\rho_1 := (\mu_{BB^T}\lambda_B L_q^2)/(c\alpha) + \sigma_G/2 + c\lambda_B^2(2-\alpha) > 0$ and $\rho^* := c\lambda_B^2(2-\alpha)$, we have:

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \frac{\rho_1}{\tau^*}\|\mathcal{L}_c^k - \mathcal{L}_c^{k+1}\|^2 + \rho^*\|w^{k+1} - w^*\|^2 \leq \max\{\frac{\rho_1}{\tau^*}, \rho^*\}\left(\mathcal{L}_c^k - \mathcal{L}_c^{k+1} + \|w^{k+1} - w^*\|^2\right). \tag{2.56}$$

Apply (2.56) to Lemma 2.4.12, we prove that the rate of convergence of the sequence $\{w^k\}_{k>N_0}$ with $N_0 \in \mathbb{N}$ is R-linear for $0 < \alpha < 1$. Combine the result with that of $1 \leq \alpha < 2$, we conclude that the rate of convergence is locally linear for $0 < \alpha < 2$. □

2.5 Applications

In this section, we apply the theoretical results developed in the last section to the IB and PF problems, which results in new solvers for each problem and improves and simplifies existing ones in terms of the convergence condition and regularization techniques. It turns out that, for the IB methods, our new solvers can achieve the same asymptotic rate of convergence to the Blahut-Arimoto (BA) typed solver, which often serves as a benchmark. In addition, the new solver, under some mild assumptions, converges independent of the choice of the trade-off parameter γ. This implies that the convergence is invariant to the potential presence of phase transitions [76] that slows down IB solvers. On the other hand, as for the PF, we develop a class of new solvers that can tackle both random and deterministic mappings, which is better than the existing greedy solvers which are limited to deterministic mappings only [3], [30]. Therefore, our new PF solvers can potentially characterize the privacy-utility trade-off better than existing ones.

2.5.1 Douglas-Rachford Splitting Based IB Solvers

Recall in section 2.1.2, we explicitly showed that the IB problem can be a special case of the proposed generalized Lagrangian (2.1). Specifically, by selecting the set of coefficients as $\rho_z := \gamma - 1, \rho_{z|x} := -\gamma, \rho_{z|y} = 1$, where $0 < \gamma < 1$ for non-trivial solutions due to the data-processing inequality [19]. It turns out that under this selection of parameters, there are

two different types of DRS solvers that corresponds to the two types of algorithms discussed in Chapter 2.2. To see this, we present the two DRS-IB solvers next. The first type of the proposed DRS-IB solver is called the "TH-DRS", which has the following formulation:

$$\begin{aligned} &p := p_z, \quad q := p_{z|x}, \quad p_{z|y} = Q_{x|y} p_{z|x}, \\ &F(p) := (\gamma - 1) H(Z), \\ &G(q) = -\gamma H(Z|X) + H(Z|Y), \\ &A = I_{N_z}, \quad B = Q_x, \end{aligned} \tag{2.57}$$

where the notation for (conditional) probability vectors and the structure of the matrices $Q_{x|y}, Q_x$ in the linear constraints are defined in (2.2). The name "TH-DRS" is chosen as such because it is first proposed in our earlier work [34] but here we simplify the design in which fewer parameters are required to decide and no additional regularization terms is imposed to assure convergence. Note that for TH-DRS, the conditional probability vector $p_{z|y}$ is treated as a equality constraint, that is, given a $p_{z|x}$ we always have $p_{z|y} = Q_{x|y} p_{z|x}$, which is in contrast to the linear penalty $p_z - Q_x p_{z|x}$ as the absolute difference decreases as the algorithm iterates. In this sense, we can think of the TH-DRS solver relaxes the marginal probability relation that p_z should have during the optimization process and finally resumes this physical meaning once converged. On the other hand, with a slight re-arrangement, we obtain the second type of the DRS-IB solvers, which is called the "MV-DRS":

$$\begin{aligned} &p := p_{z|x}, \quad q := \begin{bmatrix} p_z^T & p_{z|y}^T \end{bmatrix}^T, \\ &F(p) := -\gamma H(Z|X), \\ &G(q) := (\gamma - 1) H(Z) + H(Z|Y), \\ &A = \begin{bmatrix} Q_x^T & Q_{x|y}^T \end{bmatrix}^T, \quad B = I_{N_q}, \end{aligned} \tag{2.58}$$

where $N_q := |\mathcal{Z}| \times (|\mathcal{Y}| + 1)$. The second type is called the "MV-DRS" as this type of IB solvers can be easily generalized to multi-view IB problems that recently gain significant attention for reduced complexity over treating all views of observations as a giant view and improved the performance over single-view approaches in a variety of learning problems.

We defer the discussion of the generalization to multi-view IB problem to Chapter 3. Note that the augmented variable in MV-DRS, q is constructed by stacking the (conditional) probability vectors $p_z, p_{z|y}$ together, and therefore the total number of variables to optimize is more than that of the TH-DRS. Another observation is that since in IB the Markov chain $Y - X - Z$ is imposed on $p_{z|x}$, MV-DRS can be viewed as relaxing the Markov chain relation $p_{z|y} - Q_{x|y}p_{z|x}$ to a linear penalty constraint and retains this restriction upon converged.

After presenting the two formulations, we apply the theoretic convergence results to each formulation respectively. Recall that the results are based on different sets of assumptions, so the goal is to show that the two formulations satisfy the conditions listed in a set of assumptions. Interestingly, we find that the TH-DRS satisfies Assumption A while the MV-DRS satisfies Assumption B and therefore apply to Theorem 2.4.15 and Theorem 2.4.18 respectively under infimal measure constraints. For the first type (2.57), observe that once assumed ε_z-infimal for the primal variable p_z and $\varepsilon_{z|x}$-infimal for the augmented variable $p_{z|x}$, one of its two sub-objective function F is strongly convex while G is a combination of convex and concave functions. For G, by Lemma 2.4.4, it satisfies the definition of restricted weak convexity with respect to the matrix Q_x. As a result, we have the following convergence guarantee for the TH-DRS solver.

Theorem 2.5.1. *Suppose p_z is ε_z-infimal and $p_{z|x}$ is $\varepsilon_{z|x}$-infimal, then for $0 < \alpha \leq 2$, the IB problem formulated as in (2.57) satisfies:*

- *$F(p_z)$ is $1-\gamma$-strongly convex and $1/\varepsilon_z$-smooth.*
- *$G(p_{z|x})$ is $[(2N_zN_x\zeta)/\varepsilon_z - \gamma]$-restricted weakly convex and L_q-smooth.*

Moreover, the sequence $\{w^k\}_{k\in\mathbb{N}}$, where $w^k := (p_z^k, Q_{x|y}p_{z|x}^k, \nu^k)$ converges at linear rate locally to a stationary point when solved with the algorithm (2.4) with a penalty coefficient:

$$c > \max\left\{\frac{2N_zN_x}{\varepsilon_{z|x}}, \frac{1/\varepsilon_z + (1-\gamma)}{\alpha}\right\}.$$

Proof. Due to the ε-infimal assumptions, the Lipschitz smoothness coefficients for F and G are $L_p = 1/\varepsilon_z$ and $L_q = 1/\varepsilon_{z|x}$, respectively. Moreover, by the formulation (2.57), $F(p)$ is a scaled negative entropy function hence a strongly convex function with $\sigma_F = 1 - \gamma > 0$. As

for the function $G(q)$, since $p_{z|y} = Q_{x|y}p_{z|x}$ is a strict restriction, from Lemma 2.4.4, $G(q)$ is ω_G-restricted weakly convex w.r.t. the full row rank matrix $B = Q_x$ with the coefficient:

$$\omega_G := \frac{2N_zN_x\zeta}{\varepsilon_z} - \gamma > 0,$$

where ζ is defined as in Lemma 2.4.4. Lastly, since A is simply an identity matrix, $\lambda_A = \mu_A = 1$. By substituting the above coefficients into Lemma 2.4.5 to obtain the smallest penalty coefficient that assures convergence, it is clear that Assumption A is satisfied, and we therefore complete the proof. □

On the other hand, for the second type (2.58), upon imposing $\varepsilon_{z|x}$-infimality on the primal variables $p_{z|x}$; ε_q-infimality on the augmented variables q, one of the two sub-objective functions F is convex while the other, G is weakly convex. Hence, we can apply Theorem 2.4.18 to have the following result.

Theorem 2.5.2. *Suppose $p_z, p_{z|y}, p_{z|x}$ are $\varepsilon_z, \varepsilon_{z|y}, \varepsilon_{z|x}$-infimal, respectively, then for $0 < \alpha < 2$, the IB problem formulated in (2.58) satisfies:*

- *F is convex and $1/\varepsilon_{z|x}$-smooth*
- *G is $(2N_zN_y)/\varepsilon_{z|y}$-weakly convex and $\max\{1/\varepsilon_z, 1/\varepsilon_{z|y}\}$-smooth.*
- *The matrix $A := \begin{bmatrix} Q_x^T & Q_{x|y}^T \end{bmatrix}^T$ is full row rank.*

Moreover, the sequence $\{w^k\}_{k\in\mathbb{N}}$, where $w^k := (p_{z|x}^k, q^k, \nu^k)$ converges at linear rate to a stationary point when solved with the algorithm (2.5) with a penalty coefficient:

$$c > \frac{\alpha\sigma_G + \sqrt{\alpha^2\sigma_G^2 + 8(2-\alpha)L_q^2}}{4 - 2\alpha},$$

where $L_q := 1/\varepsilon_q$ with $\varepsilon_q := \min\{\varepsilon_z, \varepsilon_{z|y}\}$ and $\sigma_G := (2N_zN_y)/\varepsilon_{z|y}$.

Proof. Due to the ε-infimal assumptions, the Lipschitz smoothness coefficients for the functions F and G are $L_p := 1/\varepsilon_{z|x}$ and $L_q = \max\{1/\varepsilon_z, 1/\varepsilon_{z|y}\}$, respectively. Moreover, from the formulation (2.58), $F(p)$ is a negative conditional entropy which is a convex function w.r.t.

$p_{z|x}$. On the other hand, the function $G(q)$ consists of a strongly convex function $(1-\gamma)H(Z)$ w.r.t. p_z and a concave function $H(Z|Y)$ w.r.t. $p_{z|x}$. The strongly convex part does not contribute to the weak convexity of G so we can focus on $p_{z|y}$. Then since we assume $\varepsilon_{z|y}$-infimal, by Lemma 2.4.3, $H(Z|Y)$ is weakly convex with the coefficient $\sigma_G := (2N_zN_y)/\varepsilon_{z|y}$. Lastly, by construction, $B = I$, so $\mu_B = \lambda_B = \mu_{BB^T} = 1$. Substitute the coefficients mentioned above into Lemma 2.4.7 to obtain the smallest penalty coefficient that assures convergence, hence Assumption B is satisfied. We therefore complete the proof. □

From Theorem 2.5.1 and 2.5.2, it is clear that the convergence guarantee heavily depends on the infimality constraints. In literature, this assumption is commonly adopted in density or entropy estimation research [67], [68] for smoothness purposes which is aligned with our need. Another remark from the two theorems is that we can compare the two DRS-IB solvers in terms of their smallest penalty coefficient. To avoid confusion, we denote them as c^*_{TH}, c^*_{MV} respectively. The first observation is that c^*_{TH} depends on the infimality coefficient $\varepsilon_{z|x}$ whereas c^*_{MV} does not. The importance of this difference is that when $\gamma \to 1$, $p_{z|x}$ converges to deterministic mapping where the conditional probability is either 1 or 0 hence $1 \gg \varepsilon_{z|x}$, in turns results in $c^*_{TH} \gg c^*_{MV}$ since c^*_{MV} is independent of $\varepsilon_{z|x}$. The second observation is that c^*_{MV} is independent of the control parameter γ for the relevance-complexity trade-off. This implies when using MV-DRS, one can fix a penalty c when sweeping a range of γ to characterize the trade-off region. In other words, no change is needed for different choice of γ for convergence assurance. However, as mentioned earlier, MV-DRS has more parameters to optimize than that of TH-DRS. Lastly, in comparing the rate of convergence to existing IB solvers, the locally linear rates of the two DRS IB implied by Theorem 2.5.1 and 2.5.2 is the same asymptotic rate of the Blahut-Arimoto typed solver [24]. Therefore, our results are in accordance to the rate of convergence for benchmark solvers. In Chapter 2.6.1, we will evaluate the new IB solvers, and compare them to existing ones, on synthetic datasets.

2.5.2 Douglas-Rachford Splitting Based PF Solvers

In addition to IB, as demonstrated in chapter 2.1.2, our general framework (2.1) includes the PF problem as a special case by selecting the coefficients in (2.1) as:

$$\rho_z := \beta - 1, \quad \rho_{z|y} := -\beta, \quad \rho_{z|x} := 1, \tag{2.59}$$

we decompose the PF problem into a combination of a convex function $-\beta H(Z|Y)$ w.r.t. $p_{z|y}$ and $(\beta-1)H(Z)$ w.r.t. p_z. The latter sub-objective function is strongly convex if $0 < \beta < 1$. However, we empirically find that solutions for PF converge to non-trivial solutions only when $\beta > 1$. Hence, $(\beta - 1)H(Z)$ is a concave function w.r.t. p_z. Observe that by Lemma 2.4.3 the positive (conditional) entropy function is weakly convex if it is smooth, we therefore obtain the following formulation for the PF problem:

$$\begin{aligned} &p := p_{z|y}, \quad q := p_{z|x}, \quad Q_x p_{z|x} = p_z, \\ &F(p) := -\beta H(Z|Y), \\ &G(q) := (\beta - 1)H(Z) + H(Z|X), \\ &A = I_{N_z N_y}, \quad B = Q_{x|y}. \end{aligned} \tag{2.60}$$

In (2.60), linear penalty $Ap - Bq$ is simply the Markov chain relation:

$$p(z|y) = \sum_x p(z|x)p(x|y),$$

so the formulation is similar to that of the MV-DRS (2.58). However, note that the matrices for linear constraints in PF has A as the identity matrix while B is full row rank, this violates the Assumption B as for the case of MV-DRS. On the other hand, one of the two sub-objective functions F is convex, but not strongly convex while G is composed of two concave functions, and therefore not restricted-weakly convex as required in Assumption A. It turns out that, by the imposing infimality on $p_{z|x}$, G satisfies both Lipschitz continuity and smoothness and along with the fact that $p_{z|x}$ belongs to a compound simplex, we can adopt the sub-minimization path technique [45] to address the rank deficient issues of the

matrix $B = Q_x$. Following this, our finding is that upon assuming $\varepsilon_{z|x}$-infimality on $p_{z|x}$, then the proposed DRS-PF solvers satisfy the Assumption C. To show that the $G(q)$ in (2.60) for DRS-PF is weakly convex, we need the following result.

Lemma 2.5.3. *Given p_x, let $G(q)$ be defined as in (2.60). If $q = p_{z|x}$ is $\varepsilon_{z|x}$-infimal, then $G(q)$ is σ_G-weakly convex w.r.t. q, where $\sigma_G := \max\{2|\beta - 1|N_z/\varepsilon_{z|x}, 2N_zN_x/\varepsilon_{z|x}\}$.*

Proof. Since in (2.60) $G(p_{z|x}) = (\beta-1)H(Z) + H(Z|X)$, we can separate the proof into two parts. The first part is $(\beta-1)H(Z)$ and the second is $H(Z|X)$. For the first part, if $\beta \leq 1$, then the first part is a scaled negative entropy function which is $(1-\beta)$-strongly convex w.r.t. p_z and hence to $p_{z|x}$ as $p_z = Q_x p_{z|x}$ is a restriction by definition. Note that due to this restriction, $\varepsilon_z = \varepsilon_{z|x}$. To conclude the case for $\beta \leq 1$, we can simply discard the positive squared term introduced by strong convexity as a lower bound. On the other hand, if $\beta > 1$, for two distinct $p_z^m, p_z^n \in \Omega_z$, we have:

$$\begin{aligned} H(Z^m) - H(Z^n) &= \langle \nabla H(Z^n), p_z^m - p_z^n \rangle - D_{KL}(p_z^m || p_z^n) \\ &\geq \langle \nabla H(Z^n), p_z^m - p_z^n \rangle - \frac{1}{\varepsilon_z} \|p_z^m - p_z^n\|_1^2 \\ &\geq \langle \nabla H(Z^n), p_z^m - p_z^n \rangle - \frac{N_z}{\varepsilon_{z|x}} \|p_z^m - p_z^n\|_2^2, \end{aligned} \tag{2.61}$$

where the first inequality follows from reversing the Pinsker's inequality due to the $\varepsilon_{z|x}$-infimal assumption. Then for the first term in the last inequality, by the marginal relation $Q_x p_{z|x} = p_z$:

$$\langle \nabla_z H(Z^n), p_z^m - p_z^n \rangle = \langle Q_x^T \nabla_z H(Z^n), p_{z|x}^m - p_{z|x}^n \rangle = \langle \nabla_{z|x} H(Z^n), p_{z|x}^m - p_{z|x}^n \rangle,$$

where ∇_z denotes the gradient w.r.t. p_z and $\nabla_{z|x}$ w.r.t. $p_{z|x}$. For the second term in the last inequality, since $p_z = Q_x p_{z|x}$, $\|Q_x\| = 1$, we have:

$$\|p_z^m - p_z^n\|^2 \leq \|Q_x\|^2 \|p_{z|x}^m - p_{z|x}^n\|^2 = \|p_{z|x}^m - p_{z|x}^n\|^2.$$

Similarly, for $H(Z|X)$, we have:

$$\begin{aligned} H(Z^m|X) - H(Z^n|X) =& \langle \nabla_{z|x} H(Z^n|X), p^m_{z|x} - p^n_{z|x} \rangle - E_x[D_{KL}(p^m_{z|X} || p^n_{z|X})] \\ \geq& \langle \nabla H(Z^n|X), p^m_{z|x} - p^n_{z|x} \rangle - \frac{N_z N_x}{\varepsilon_{z|x}} \|p^m_{z|x} - p^n_{z|x}\|^2. \end{aligned}$$

Combining the two results, pre-multiplying $|\beta - 1|$ to that of $H(Z)$, we conclude that $G(p_{z|x})$ is σ_G-weakly convex w.r.t. $p_{z|x}$, where $\sigma_G := \max\{2|\beta - 1|N_z/\varepsilon_{z|x}, 2N_x N_z/\varepsilon_{z|x}\}$. □

Then since Assumption C is satisfied for DRS-PF, we can apply Theorem 2.4.20 to provide convergence guarantee for the new class of solvers.

Theorem 2.5.4. *Suppose $p_{z|y}, p_{z|x}$ are $\varepsilon_{z|y}, \varepsilon_{z|x}$-infimal respectively, then for $0 < \alpha < 2$, the PF problem formulated in (2.60) satisfies:*

- *F is convex and $1/\varepsilon_{z|y}$-smooth.*
- *G is $[2N_z(|\beta - 1| + N_x)]/\varepsilon_{z|x}$-weakly convex, $1/\varepsilon_{z|x}$-smooth and $2|\log \varepsilon_{z|x}|$-Lipschitz continuous.*
- *The matrix $B := Q_{x|y}$ is full row rank.*

Moreover, the sequence $\{w^k\}_{k\in\mathbb{N}}$, where $w^k := (p^k_{z|y}, p^k_{z|x}, \nu^k)$ converges at linear rate locally to a stationary point when solved with the algorithm (2.5) with a penalty coefficient:

$$c > M_q \left[\frac{M_q \alpha \sigma_G + \sqrt{M_q^2 \alpha^2 \sigma_G^2 + 8(2-\alpha) L_q^2 \lambda_B^2 / \mu_{BB^T}}}{4 - 2\alpha} \right],$$

where $\sigma_G := [2N_z(|\beta - 1| + N_x)]/\varepsilon_{z|x}$, $M_q := 2|\log \varepsilon_{z|x}|$, and $L_q := 1/\varepsilon_{z|x}$.

Proof. By assumption, $p_{z|y}$ is $\varepsilon_{z|y}$-infimal, and $p_{z|x}$ is $\varepsilon_{z|x}$-infimal. Hence, by Corollary 2.4.2, the Lipschitz smoothness coefficients for the functions F and G are $L_p := 1/\varepsilon_{z|y}$ and $L_q = 1/\varepsilon_{z|x}$, respectively. Moreover, from the formulation (2.60), $F(p) = -H(Z|Y)$ is a convex function w.r.t. $p := p_{z|y}$ as shown in Corollary 2.4.2. On the other hand, for the function $G(q)$, by Lemma 2.5.3, G is $2N_z[|\beta - 1| + N_x]/\varepsilon_{z|x}$-weakly convex w.r.t. $q := p_{z|x}$. The $\varepsilon_{z|x}$-infimal assumption implies the Lipschitz continuity of G, which can be shown by combining

Lemma 2.4.1 and Corollary 2.4.2. In turns, since the q-update (2.5c) is equivalent to the Lipschitz continuous function $\Phi(\mu) := \arg\min_{q\in\Omega_q} G(q) + c/2\|Bq - \mu\|^2$, due to the fact that $Bq = \hat{p}_{z|y}$ is bounded, there exists a sub-minimization path [45] such that the following holds:

$$\|q^m - q^n\| \leq \|\Phi\left(Bq^m\right) - \Phi\left(Bq^n\right)\| \leq M_q\|Bq^m - Bq^n\|,$$

where $M_q := 2|\log \varepsilon_{z|x}|$ denotes the Lipschitz continuity coefficient of G, and hence of Φ. As for the linear constraints, since the matrix $A = I$, we have $\mu_A = \lambda_A = \lambda_{AA^T} = 1$ whereas $B = Q_{x|y}$ as constructed in (2.60). Note that B is full row rank since each row corresponds to a conditional prior probability and if there are identical rows, we can simply eliminate the duplicate rows as they represent the same conditional distribution of the observations. As a result, we have $\lambda_{BB^T}, \mu_{BB^T}$ as the largest and smallest eigenvalues of $Q_{x|y}Q_{x|y}^T$. Substitute the coefficients $M_q, \sigma_G, \lambda_B, \mu_{BB^T}$ into Lemma 2.4.7 to obtain the smallest penalty coefficient that assures convergence, and we conclude that the Assumption C is satisfied, which completes the proof. □

As a remark, from Theorem 2.5.4, the proposed DRS-PF solvers converges linearly to a stationary point, this result does not restrict the elements of the primal variables $p_{z|x}$ to be either 1 or 0, which is often imposed for existing clustering based greedy algorithms [3], [30]. Our new PF solvers are therefore capable of handling both deterministic and random mappings so potentially can explore the privacy-utility trade-off [3] in PF research better than existing ones. We empirically evaluate the proposed PF-DRS solvers and compare them with existing ones in Chapter 2.6.2.

2.6 Evaluation

In this section, we implement the proposed algorithms for IB and PF and evaluate them on both synthetic and real-world datasets. In implementing the algorithms, we adopt gradient descent to update the primal variables $p_z, p_{z|x}, p_{z|y}$. It is worth noting that, to make the

gradient updates on the primal variables be valid probabilities, it turns out that the updated becomes the mean-subtracted gradients, for example, when updating $p^k_{z|x}$:

$$p^{k+1}_{z|x} := p^k_{z|x} - \epsilon^k_{z|x} \nabla_{z|x} \bar{\mathcal{L}}^k_c,$$

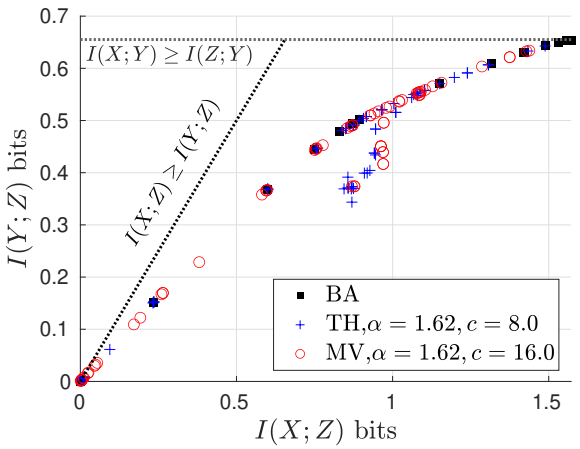
where $\nabla_{z|x} \bar{\mathcal{L}}^k_c := \nabla_{z|x} \mathcal{L}^k_c - \frac{1}{N_z} \sum_z \nabla_{z|x} \mathcal{L}^k_c$, with a stepsize $\epsilon_{z|x}$ chosen to assure that the updates remain valid (conditional) probabilities. Similarly, the other primal variables $p_z, p_{z|y}$ are updated in mean-subtracted gradients, but with respect to the parts in the augmented Lagrangian they are involved in. We start with synthetic data to evaluate the proposed DRS-IB solvers. The joint probability $p(X, Y)$, with dimensions $N_x = N_y = N_z = 3$, is as follows.

$$p(Y|X) = \begin{bmatrix} 0.90 & 0.08 & 0.40 \\ 0.025 & 0.82 & 0.05 \\ 0.075 & 0.10 & 0.55 \end{bmatrix}, \quad p(X) = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}^T.$$

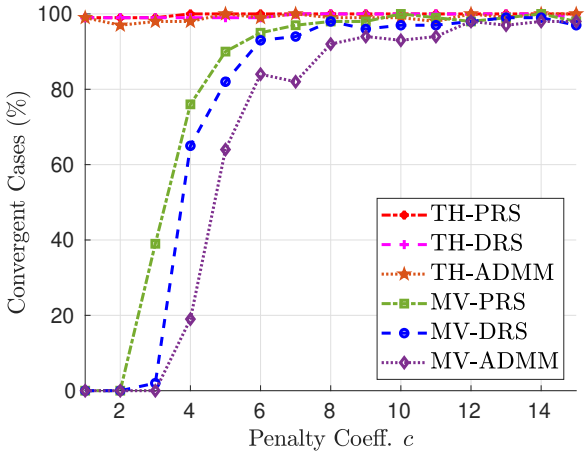
2.6.1 Evaluation on a Synthetic IB Problem

As reviewed in Chapter 1.2.1, the performance of solutions is evaluated on the information plane [5], [6], which serves as a reference to find the maximum $I(Y; Z)$ one can achieve given a fixed $I(X; Z)$. By convention, we compare the solutions obtained from the proposed two solvers, e.g. DRS-TH and DRS-MV, to those obtained from the BA-typed algorithm (as a benchmark). In the following figures, the BA-typed solver will be denoted as *BA*. For the proposed approaches, we denote the algorithm (2.4) as *TH* whereas (2.5) as *MV*. For the relaxation parameters α, we choose $\alpha \in \{1.0, 1.618, 2.0\}$ corresponding to *ADMM*, *DRS* and *PRS*. The choice of $= \frac{1+\sqrt{5}}{2} \approx 1.618$ for the *DRS* is simply inspired by the known result that this is the maximum possible relaxation parameter for DRS in convex settings [72]. Notably, our earlier work, the two-block ADMM-IB solver [34], corresponds to the algorithm (2.4) with $\alpha = 1$ (*TH-ADMM*).

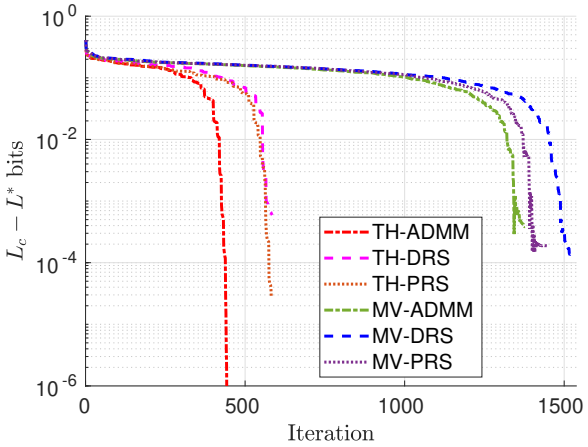
For each algorithm, we randomly initialize $p_{z|x}, p_z$ and run the algorithm until it terminates. This procedure is repeated for 100 times for each $\beta \in [1.0, 10.0]$. For the proposed



(a) Information plane



(b) Percentage of Convergent Cases, $\gamma = 0.286$



(c) $\mathcal{L}_c$ Decrease $\gamma = 0.154$

Figure 2.1. Simulation results of IB on a synthetic joint probability $p(X, Y)$ with $N_z = N_x = N_y = 3$

methods, the stopping criterion is either the residuals, $\left\|p_z^k - Q_x p_{z|x}^k\right\|_1^2$ for the algorithm (2.4) or $\left\|Ap^k - q^k\right\|_1^2$ for (2.5), is less than 2×10^{-6}. This will be labeled as a convergent case. Otherwise, if a maximum number of iterations is reached, then we label it as a divergent case. Our simulation results are presented in Figure 2.1. In Figure 2.1a, for each γ, we collect the results from 100 runs including both convergent and divergent cases, but only the convergent cases are scattered. In Figure 2.1b and 2.1c, *TH* is referred to the algorithm (2.4) whereas *MV* for (2.5). For each algorithm, we have three sub-labels *PRS, DRS, ADMM*, corresponding to $\alpha = \{2.0, 1.62, 1.0\}$. In Figure 2.1c,$\mathcal{L}^* = -0.411$ bits and the penalty coefficient for *TH* is $c_{TH} = 8.0$ while $c_{MV} = 16.0$.

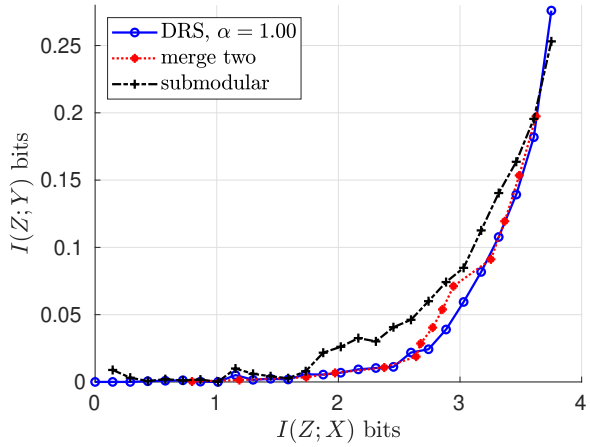
In Figure 2.1a, we first compare the information plane characterized by the proposed splitting methods based solvers to that obtained from *BA* by mappings each $[I(X;Z), I(Y;Z)]$ pair. Clearly, the correctness of the two splitting methods based algorithms can be confirmed since the solutions obtained from them achieve comparable performance to that obtained from the BA-typed solver. Note that in Figure 2.1a, there are local minima due to the intrinsic non-convexity of the IB problems.

Then we focus on the splitting methods and the effect of the relaxation parameter α. Observe that, from Theorem 2.5.1 and 2.5.2 the smallest penalty coefficient c^* that assures convergence is related to α, we therefore use the same synthetic joint probability matrix to further evaluate the percentage of convergent cases empirically for the two types of solvers. In Figure 2.1b, we fix a trade-off parameter $\gamma = 0.286$ and vary the penalty coefficient $c \in [1, 15]$. In Figure 2.1b, we show the percentage of convergent cases versus c. Comparing the convergence performance of *TH* and *MV*, we find that the percentage of convergent cases is almost 100% even with $c = 1$ whereas the smallest penalty coefficient that assures high convergent percentage for *MV* is $c^*_{MV} \approx 12.0$. This is because there are more variables to optimize for *MV* over *TH*. Interestingly, *MV-PRS* ($\alpha = 2.0$) has highest c^*, and since $\alpha \approx 1.618$ for the *MV-DRS*, our simulation results show that the the relaxation step improves the convergence of the splitting methods in IB in the sense that a smaller c^* is needed for higher percentage of convergent cases.

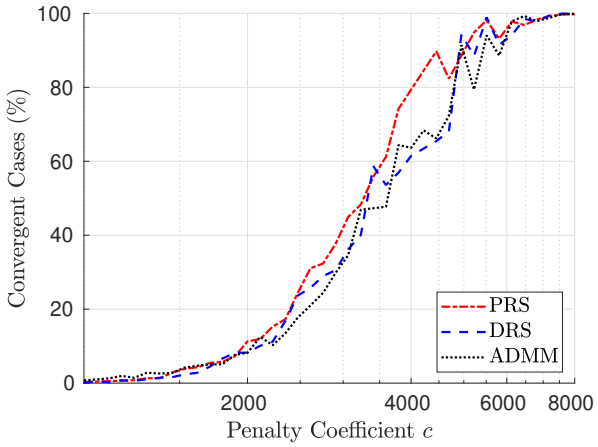
Lastly, as our theoretic convergence analysis implies that the corresponding rates of convergence of the two algorithms are linear, we examine the progression of the loss, i.e., the augmented Lagrangian, over iteration. The result is shown in Figure 2.1c, where we fix a $\gamma = 0.154$. The minimum loss $\mathcal{L}^* \approx -0.411$ (bits) for the augmented Lagrangian (2.3), which is found after the six types of solvers converged. For *TH*, we fix the penalty $c_{\text{TH}} = 8.0$ whereas $c_{\text{MV}} = 16.0$ for *MV* instead. In consistent with in the results for the percentage of convergent cases, since *TH* has less variables to optimize, fewer iterations are performed to reach convergence $\mathcal{L}^*$. Then, focusing on the three types of solvers in either *TH* or *MV* in Figure 2.1c, we find that *ADMM* ($\alpha = 1$) requires less number of iteration to converge in both *TH* and *MV*, so while α improves convergence, the rate is slower for a higher α. Note that since we implement the algorithms with gradient descent, the rate of convergence also depends on the selection of step size [21], [72], [77]. Based on our theoretic results, e.g., Lemma 2.4.14 and 2.4.17, we expect for a sufficient large c, the convergence is locally linear. Clearly this is the case as Figure 2.1c shows. For *TH*, after about 500 iterations, it converges linearly toward $\mathcal{L}^*$ and similar behavior can be observed for *MV*, which happens after around 1250 iterations instead. The empirical results therefore confirm our theoretical analysis on the convergence and the locally linear rates.

2.6.2 Experiments for PF Solvers

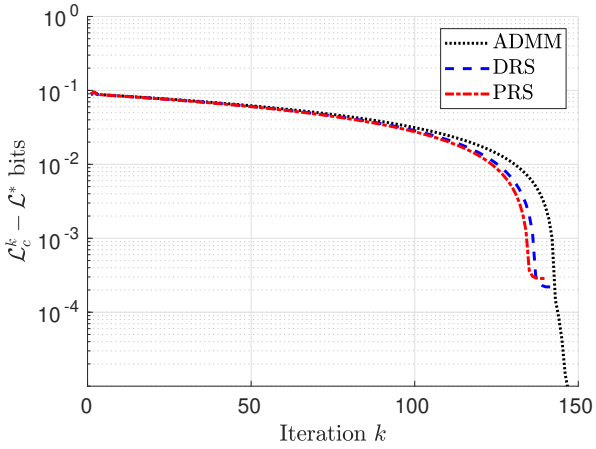
For the proposed new splitting methods based PF solvers, we evaluate the algorithm on real-world data. The dataset is named "Heart failure clinical records Data Set" [78] from the UCI Machine Learning Repository [79]. It has 299 instances with 13 attributes. Among which, we select 6 attributes including: "anaemia," "high blood pressure," "diabetes," "sex," "smoking" and "death". All selected attributes are binary. In preparing the joint probability, we let $Y :=$ {"sex","death"} and the rest be X, which gives $|Y| = 4, |X| = 16$. The joint probability is formed by counting the 299 instances with respect to each (Y, X) pair. Afterward we add 10^{-3} to each entry of the counted results to avoid $p(x, y) = 0$. The sensitive information Y, in PF context, consists of 2 binary labels {"death","sex"}. In Figure 2.2a, we collect all the convergent pairs of $[I(Z;X), I(Z;Y)]$ with varying $\beta \in [1.0, 10.0]$, $2 \leq |\mathcal{Z}| \leq 16$



(a) Information Plane



(b) Percentage of Convergent Cases, $\beta = 3.5$



(c) $\mathcal{L}_c$ Decrease, $\beta = 3.5$

Figure 2.2. Experiment results for PF on UCI-heart failure clinical records dataset

and $c = 7000$. Then we show the minimum $I(Z;Y)$ for given a $I(Z;X)$. In Figure 2.2b, *PRS*,*DRS*,*ADMM* corresponds to $\alpha = \{2, 1.62, 1\}$ respectively. For Figure 2.2b and 2.2c, we fix $|\mathcal{Z}| = 16$ and set the penalty coefficient to 7000. We further let the three splitting methods start from the same initial point and the minimum function value that is achieved by the three methods is $\hat{\mathcal{L}}_c^* = -2.46$.

We implement the algorithm (2.5) based on (2.60). For the relaxation parameter, we choose $\alpha = \{1.00, 1.618, 2.00\}$ correspond to *ADMM*, *DRS*, *PRS* respectively. For the compared methods, we select two existing PF solvers that are clustering-based algorithms, the first one [3] greedily merging two $x_1, x_2 \in \mathcal{X}$ to form a new cluster Z_{new} that maximize $I(Z_{old};Y) - I(Z_{new})$. We denote this solver as *merge two*. Another solver we compared with is [30], where it relaxes the sub-optimal pairwise merging approach, by using submodular-supermodular optimization approach to consider subsets of more than two elements, however it is limited to deterministic mappings as *merge two*. We denote the second greedy algorithm as *submodular*.

Our experimental results are shown in Figure 2.2. We start with evaluating the solution obtained from the proposed DRS solvers to that from the two compared solvers. Note that, since other choice of the relaxation parameter α gives similar results when scattered on the information plane, for clarity purposes, we only show $\alpha = 1.00$. In Figure 2.2a the three methods give comparable performance on the information plane which verified the correctness of our implementation. But different from the two clustering based approaches, the proposed DRS based PF solvers are not limited to deterministic mappings, and therefore can explore the area between the consecutive points obtained from the compared methods. This can be observed from the regions $I(X;Z) < 1$ compared to *merge two* and $I(X;Z) < 3.5$ compared to *submodular* in Figure 2.2a.

To explore the information plane with the proposed DRS solver as in Figure 2.2a, we vary the trade-off parameter $\beta \in [1, 10]$ with random initialization for multiple times and collect the solutions that converged. Note that before starting the proposed PF solvers, the penalty coefficient c and α need to be determined. From Theorem 2.5.4, there exists a smallest penalty coefficient c that assures convergence and can be determined if the conditions listed are satisfied, but in practice, one can increase c until a desired convergent percentage is

reached. In Figure 2.2b we fix $\beta = 3.5$ and sweep $c \in [1000, 8000]$. For each c, we randomly initialize $p_{z|x}$ and run the algorithm until a termination condition is reached. This procedure is repeated for 2000 times for each method. Then we report the average percentage of convergent cases. From the figure, observe that for a given percentage, *PRS* has a smallest penalty coefficient than *DRS*, *ADMM*, this again demonstrate the relaxation α improves convergence and is consistent with the insight we have for the IB problem in the last part.

Lastly, we evaluate the rate of convergence of the proposed DRS solvers for PF. In Figure 2.2c we fix an initial point for the three solvers and set the penalty coefficient $c = 7 \times 10^3$. The minimum value of the augmented Lagrangian among the three is $\hat{\mathcal{L}}_c^* = -2.46$. Recall that by Theorem 2.4.20 we claim the rate of convergence of the DRS solver is locally linear, the results shown in Figure 2.2c confirm our claim. Observe that around the iteration count $k = 100$, the variables are getting closer toward a local minimum and converges linearly fast afterward. Interestingly, contrary to the observation in the IB evaluation, in PF, *PRS* reaches converges to the local minimum faster than *DRS* and *ADMM*, so the relaxation parameter α in this experiment improves both convergence and its rate. Finally, we note that there are implementation issues that affect the evaluation or the selection of c and α. For example, the step size selection schemes [21] or acceleration techniques for gradient descent algorithms [72], we leave these practical issues for future exploration.

3. MULTI-MODAL IB AND PF REPRESENTATION LEARNING

In this chapter, we generalize the single-modal IB and PF to multi-modal scenarios. For clarity, we will focus on multi-view IB (MvIB) first and generalize to the multi-source PF (MsPF) in section 3.2.3. In multi-view learning literature, the basic assumption is the *conditional independence*, that is, the joint probability of observations of all views $\{X\}_{\mathrm{i}=1}^{V}$ generated from a target variable Y has the relation: $p(y, \{x\}_{\mathrm{i}=1}^{V}) = p(y)\Pi_{\mathrm{i}=1}^{V} p(x^{(\mathrm{i})}|y)$. Based on this condition, we adopt an information-theoretic approach to formulate the MvIB and MsPF.

3.1 Main Results

The novelty of this part is our new information-theoretic formulations of the MvIB and MsPF problems. This is in sharp contrast to existing heuristic-based forms as (3.1) and (3.16) reduce to single-view IB and PF respectively. This implies that the solutions under the proposed formulations can map to the information plane in IB and PF to have performance references. The new formulations are devoted to solving the two challenges in multi-modal learning: the performance-complexity trade-off and the representation overlap. For the first challenge, consider MvIB, the naïve approach is to merge all views into a giant view which loses no information, but the observation dimension grows exponentially w.r.t. the number of views. On the other hand, due to conditional independence, one can also pick the view that outperforms the others which will have minimum complexity but exploit no performance gain from multi-view observations. Therefore, the interpretation of the performance-complexity trade-off is finding representations that avoid the exponential growth of the dimensions but enjoy the performance improvement from multi-view observations.

As for the second challenge, consider a scenario where the observations consist of two forms: images and videos, in this case, as videos can be thought of as a time sequence of images, the two clearly have abundance of representation overlap. In contrast, suppose the two forms are: images and audio, then the representation overlap is limited as it is not easy

to learn color or shape from sounds. However, even with limited overlap, learning with both views clearly provides a more complete representation than learning through each view alone. In short, the challenge is learning representations from heterogeneous forms of observations but whether the overlap is abundant or limited, enjoying performance gain from multi-view learning is possible.

To jointly address the two challenges, we propose two new formulations that both significantly reduce the complexity growth as compared to the merge-view approach inspired by the two extremes of the abundance of the representation overlap. When there is substantial overlap, we propose the consensus-complement formulation (Section 3.2.1) whereas when there is limited overlap, we propose the incremental update (Section 3.2.2). The novelty is that through the associated restrictions imposed to reduce complexity growth in the two formulations, the obtained Lagrangians (3.2) and (3.13), are both generalizations of the single-view Markovian Lagrangian (2.1) studied in Chapter 2 and hence we propose solving the two MvIB Lagrangians through the splitting methods.

Our key observation is that we can generalize the single-view approach introduced in the last chapter to study the two proposed formulations. First, for the consensus-complement form, we propose solving the overall objective function in two steps where in the first step a consensus representation is learned then it becomes the side-information for learning view-specific, complement representation independently across views. Through this simplification, the extra sub-objectives in consensus Lagrangian (3.6) are all convex as compared to (2.1) whereas the complement Lagrangian (3.6) have the same three-block set-up but each sub-objective is conditioned on the consensus representation. It turns out that by Bayes' rule and Markov relation, the additional conditional probabilities related to the consensus representation can be treated as equivalent priors and recovers the single-view Markovian Lagrangian. Overall, for the consensus-complement formulation, we can formulate both steps to consensus augmented Lagrangians, (3.9) and (3.10) respectively, and solve it with consensus ADMM (3.11) as the complement step is, by our formulation, is simply a special case where there is only one view. Remarkably, as the extra sub-objective functions in consensus augmented Lagrangian are all convex, we can easily extend the convergence results in last chapter to prove locally linear rate of convergence for the proposed solver (Theorem 3.3.7).

Second, for the incremental update formulation, the key idea is that if the overlap is limited, then accumulating the learned representation view-by-view can be significantly simpler than forming a consensus representation first. By imposing the restrictions to achieve this goal (Section 3.2.2), the overall objective function is decomposed into a level-structured sum of Lagrangians (3.14). By construction, the lower-level representations are accumulated and provided as side-information for current-level representation learning. Therefore, each level of the incremental update Lagrangian consists of three sub-objectives and hence can be re-formulated to augmented Lagrangian as in the complement step of the last form. The only difference is that before moving on to the next level, the accumulated representation, or the equivalent prior probabilities, need to be updated (3.15). Since the algorithm for each level of the incremental update is again a special case of the consensus ADMM, the convergence analysis extends to this form as well.

In evaluation (Section 3.4), defining the complexity as the asymptotic growth of the dimensions of the mappings from observations to representations, we show that the complexity for the consensus-complement form is linear w.r.t. the number of views while it is exponential with the factor of the dimension of the representations for the incremental update. Notably, compared to the merge-view approach whose complexity is exponential with the factor of the dimension of the observations, we show that both the proposed formulations significantly reduce the complexity compared to the merge-view approach as in general a low-dimensional representation is preferred. Then, in evaluating the performance loss due to the associated restrictions for the two proposed formulations, we show that both approaches suffer from slight performance loss as compared to the merge-view approach (served as the optimal performance reference) but the two both outperform the state-of-the-art DNN based solver in a wide range of model configurations.

3.2 Information-Theoretic Formulation of Multi-View IB

The goal of MvIB is to design a set of representations $\{Z\}$ with access to individual view-specific observation $X^{(\mathrm{i})}, \forall \mathrm{i} \in [V]$ that maximize the relevance to a target variable Y, measured in mutual information $I(Y;\{Z\})$ while minimizing the compression rate to all

views of observations $I(\{X^{(\mathrm{i})}\}_{\mathrm{i}=1}^{V};\{Z\})$. The objective can be expressed as the following Lagrangian:

$$\mathcal{L}_M := \gamma I(\{X\}_{\mathrm{i}=1}^{V};\{Z\}) - I(Y;\{Z\}), \tag{3.1}$$

where $\gamma > 0$ is a fixed constant serves as the trade-off parameter between compression of each $X^{(\mathrm{i})}$ and the relevance to Y. Note that if we design the representation set to have a single element $\{Z\} := Z^*$, then (3.1) reduces to the single-modal case. Intuitively, one can combine the observations from all views into a giant single view of observations, but the dimension scales in multiplicative manner, that is, $|\mathcal{X}^*| := \Pi_{\mathrm{i}=1}^{V}|\mathcal{X}^{(\mathrm{i})}|$. The exponential growth of the complexity prevents this approach from solving practical large-scale views problems. In the following, we propose two forms of MvIB that significantly reduce the complexity over the single-modal approach but achieve comparable performance to it.

3.2.1 MvIB: Consensus-Complement Form

Inspired by the co-training methods in multi-view literature [80], in the first form, we design the set of latent representations $\{Z\}$ to have a consensus representation, denoted as Z_c and view-specific complement components, denoted as $Z_{\mathrm{e}}^{(\mathrm{i})}, \forall \mathrm{i} \in [V]$. Then, by the chain rule of mutual information, the Lagrangian of (3.1) becomes the following:

$$\mathcal{L}_{\text{con}} = \gamma I(Z_c;\{X\}) - I(Z_c;Y) + \sum_{\mathrm{i}=1}^{V} \gamma I(Z_{\mathrm{e}}^{(\mathrm{i})};\{X\}|Z_c,\{Z_{\mathrm{e}}\}_{\mathrm{i}-1}) - I(Y;Z_{\mathrm{e}}^{(\mathrm{i})}|Z_c,\{Z_{\mathrm{e}}\}_{\mathrm{i}-1}), \tag{3.2}$$

where the sequence $\{Z_{\mathrm{e}}\}_{\mathrm{j}} := \{Z_{\mathrm{e}}^{(1)},\cdots,Z_{\mathrm{e}}^{(\mathrm{j})}\}$ is defined to represent the accumulated complement views. To further simplify the above formulation, the representation set is subjected to the following conditions (similar to [9], [80]):

- There always exist constants $\kappa_{\mathrm{i}}, \forall \mathrm{i} \in [V]$, independent of the observations $\{X\}$ such that $\kappa_{\mathrm{i}} I(Z_c;X^{(\mathrm{i})}) = I(Z_c;X^{(\mathrm{i})}|\{X\}_{\mathrm{i}-1})$.
- $Y \to X^{(\mathrm{i})} \to Z_{\mathrm{e}}^{(\mathrm{i})} \leftarrow Z_c$ forms a Markov chain. That is, Z_c is the side information for $Z_{\mathrm{e}}^{(\mathrm{i})}$.
- For each view $\mathrm{i} \in [V]$, given the consensus Z_c, $\{Z_{\mathrm{e}}\}$ are independent.

Under these constraints, we can then rewrite (3.2) as a superposition of two parts: $\mathcal{L} := \bar{\mathcal{L}} + \sum_{\mathrm{i}=1}^{V} \mathcal{L}_{\mathrm{e}}^{(\mathrm{i})}$, where the first component $\bar{\mathcal{L}}$ is defined as the multi-view consensus IB Lagrangian:

$$\bar{\mathcal{L}} := \sum_{\mathrm{i}=1}^{V} \gamma_{\mathrm{i}} I(Z_c; X^{(\mathrm{i})}) - I(Z_c; Y), \tag{3.3}$$

and the second consists of V terms with each one represents a complement sub-objective for each individual view:

$$\mathcal{L}_{\mathrm{e}}^{(\mathrm{i})} := \gamma I(Z_{\mathrm{e}}^{(\mathrm{i})}; X^{(\mathrm{i})} | Z_c) - I(Z_{\mathrm{e}}^{(\mathrm{i})}; Y | Z_c), \forall \mathrm{i} \in [V]. \tag{3.4}$$

With these additional conditions, we can recast $\bar{\mathcal{L}}$ in (3.3) as:

$$\begin{aligned} \bar{\mathcal{L}} : &= -\sum_{\mathrm{i}=1}^{V} \gamma_{\mathrm{i}} H(Z|X^{(\mathrm{i})}) + \left(-1 + \sum_{\mathrm{i}=1}^{V} \gamma_{\mathrm{i}}\right) H(Z) + H(Z|Y) \\ &= \sum_{\mathrm{i}=1}^{V} F_{\mathrm{i}}(p_{z|x,\mathrm{i}}) + G(p_z, p_{z|y}). \end{aligned} \tag{3.5}$$

Similarly, we rewrite (3.4), $\forall \mathrm{i} \in [V]$, as:

$$\mathcal{L}_{\mathrm{e}}^{(\mathrm{i})} = -\gamma H(Z_{\mathrm{e}}^{(\mathrm{i})} | Z_c, X^{(\mathrm{i})}) + (\gamma - 1) H(Z_{\mathrm{e}}^{(\mathrm{i})} | Z_c) + H(Z_{\mathrm{e}}^{(\mathrm{i})} | Z_c, Y). \tag{3.6}$$

By representing the discrete (conditional) probabilities as vectors/tensors, we can solve (3.5) and (3.6) with augmented Lagrangian methods. To show this, we define the following vectors:

$$\begin{aligned} p_{z|x,\mathrm{i}} &:= \left[p(z_1|x_1^{(\mathrm{i})}) \cdots p(z_1|x_{N_{\mathrm{i}}}^{(\mathrm{i})}) \cdots p(z_L|x_{N_{\mathrm{i}}}^{(\mathrm{i})})\right]^T, \\ p_z &:= \begin{bmatrix} p(z_1) & \cdots p(z_L) \end{bmatrix}^T, \\ p_{z|y} &:= \left[p(z_1|y_1) \cdots p(z_1|y_T) \cdots p(z_L|y_K)\right]^T. \end{aligned} \tag{3.7}$$

where $N_{\mathrm{i}} := |\mathcal{X}^{(\mathrm{i})}|, \forall \mathrm{i} \in [V], L := |\mathcal{Z}|, K := |\mathcal{Y}|$. For clarity of expression, we denote the primal variables for each individual view as $p_{z|x,\mathrm{i}} := p_{\mathrm{i}}$, and cascade the augmented variables

into a single expression $q := \begin{bmatrix} p_{z_c}^T & p_{z_c|y}^T \end{bmatrix}^T$. On the other hand, for the complement part, we define the following tensors:

$$\begin{aligned} \boldsymbol{\pi}_{x,\mathrm{i}}[m,n,r] &:= P(Z_{\mathrm{e}}^{(\mathrm{i})} = z_{\mathrm{e},m}^{(\mathrm{i})} | Z_c = z_{c,n}, X^{(\mathrm{i})} = x_r^{(\mathrm{i})}), \\ \boldsymbol{\pi}_{y,\mathrm{i}}[m,n,r] &:= P(Z_{\mathrm{e}}^{(\mathrm{i})} = z_{\mathrm{e},m}^{(\mathrm{i})} | Z_c = z_{c,n}, Y = y_r), \\ \boldsymbol{\pi}_{z,\mathrm{i}}[m,n] &:= P(Z_{\mathrm{e}}^{(\mathrm{i})} = z_{\mathrm{e},m}^{(\mathrm{i})} | Z_c = z_{c,n}). \end{aligned} \tag{3.8}$$

With the above definition, we present the consensus-complement MvIB augmented Lagrangian as follows. For the consensus part:

$$\bar{\mathcal{L}}_c(\{p_{\mathrm{i}}\}_{\mathrm{i}=1}^V, q, \{\nu_{\mathrm{i}}\}_{\mathrm{i}=1}^V) = \sum_{\mathrm{i}=1}^V \left[F_{\mathrm{i}}(p_{\mathrm{i}}) + \langle \nu_{\mathrm{i}}, A_{\mathrm{i}} p_{\mathrm{i}} - q \rangle + \frac{c}{2} \| A_{\mathrm{i}} p_{\mathrm{i}} - q \|^2 \right] + G(q), \tag{3.9}$$

where $\|\cdot\|$ is in 2-norm, $c > 0$ is the penalty coefficient and the linear penalty $A_{\mathrm{i}} p_{\mathrm{i}} - q$ for each view $\mathrm{i} \in [V]$ encourages the variables q and each p_{i} to satisfy the marginal probability and Markov chain conditions. Specifically, let $\otimes$ denote the Kronecker product, $A_{x,\mathrm{i}} := I \otimes p_{x^{(\mathrm{i})}}^T, A_{x|y}^{(\mathrm{i})} := I \otimes P_{x^{(\mathrm{i})}|y}^T$ where $P_{x^{(\mathrm{i})}|y}$ is the matrix form of the conditional distribution $p(x^{(\mathrm{i})}|y)$ with each entry (m,n) equals to $p(x_m^{(\mathrm{i})}|y_n)$ and $A_{\mathrm{i}} := \begin{bmatrix} A_{x,\mathrm{i}}^T & A_{x|y}^{(\mathrm{i})} \end{bmatrix}^T$. As for the complement part, let

$$\begin{aligned} F_{\mathrm{e,i}} &= -\gamma H(Z_{\mathrm{e}}^{(\mathrm{i})} | Z_c, X^{(\mathrm{i})}), \\ G_{\mathrm{e,i}} &= (\gamma - 1) H(Z_{\mathrm{e}}^{(\mathrm{i})}) + H(Z_{\mathrm{e}}^{(\mathrm{i})} | Z_c, Y). \end{aligned}$$

For each realization $(z_{\mathrm{e}}, z_c) \in \mathcal{Z}_{\mathrm{e}} \times \mathcal{Z}_c$, define the cascaded vector:

$$\boldsymbol{\pi}_{q,\mathrm{i}}[l,m,:] := \begin{bmatrix} \boldsymbol{\pi}_{z,\mathrm{i}}[l,m]^T & \boldsymbol{\pi}_{y,\mathrm{i}}[l,m,:]^T \end{bmatrix}^T, \quad \forall (l,m) \in ([|\mathcal{Z}_{\mathrm{e}}|], [|\mathcal{Z}_c|]),$$

where the notation : denotes all the entries along the indicated axis. Using the definitions, we can recover the linear expression of the penalty term, to see this, by Bayes' rule and the Markov relation, we have:

$$p(z_{\mathrm{e}}^{(\mathrm{i})} | z_c, y) = \frac{\sum_{X^{(\mathrm{i})}} p(z_{\mathrm{e}}^{(\mathrm{i})}, z_c | x^{(\mathrm{i})}) p(x^{(\mathrm{i})} | y)}{p(z_c | y)}.$$

This implies that we can treat the denominator as an additional prior term for each $z_c \in \mathcal{Z}_c$ because Z_c is obtained in advance of the complement step. Based on this idea, we express the linear penalty for each complement step as:

$$A_{\mathrm{e}}^{(\mathrm{i})}\mathrm{vec}(\boldsymbol{\pi}_{x,\mathrm{i}}) := \begin{bmatrix} \Lambda_{z_c|y}^{-1}(1)Q_{x|y}^{(\mathrm{i})} & 0 & \cdots \\ 0 & \ddots & 0 \\ \vdots & 0 & \Lambda_{z_c|y}^{-1}(|\mathcal{Z}_c|)Q_{x|y}^{(\mathrm{i})} \end{bmatrix} \begin{bmatrix} \boldsymbol{\pi}_x^{(\mathrm{i})}(z_c = 1) \\ \vdots \\ \boldsymbol{\pi}_x^{(\mathrm{i})}(z_c = |\mathcal{Z}_c|) \end{bmatrix},$$

where $\mathrm{vec}(\cdot)$ denotes the vectorization operator of a tensor, $\Lambda_{z_c|y}(t)$ transforms a vector to a matrix by mapping its entries to the diagonal and the inner parentheses indicate the index $t \in [|\mathcal{Z}_c|]$. Using the definitions, the augmented Lagrangian for the i-th view's complement step can be expressed as:

$$\mathcal{L}_{\mathrm{e},c}^{(\mathrm{i})} = F_{\mathrm{e,i}}(\boldsymbol{\pi}_{x,\mathrm{i}}) + G_{\mathrm{e,i}}(\boldsymbol{\pi}_{q,\mathrm{i}}) + \langle \mu_{\mathrm{i}}, A_{\mathrm{e}}^{(\mathrm{i})}\mathrm{vec}(\boldsymbol{\pi}_{x,\mathrm{i}}) - \mathrm{vec}(\boldsymbol{\pi}_{q,\mathrm{i}}) \rangle + \frac{c}{2}\|A_{\mathrm{e}}\mathrm{vec}(\boldsymbol{\pi}_{x,\mathrm{i}}) - \mathrm{vec}(\boldsymbol{\pi}_{q,\mathrm{i}})\|^2, \quad (3.10)$$

where $c > 0$ is the penalty coefficient.

Next, we propose a two-step algorithm to solve (3.2) and describe the steps as follows. The first step is solving (3.9) through the following consensus ADMM algorithm. $\forall \mathrm{i} \in [V]$:

$$p_{\mathrm{i}}^{k+1} := \underset{p_{\mathrm{i}} \in \Omega_{\mathrm{i}}}{\arg\min} \quad \bar{\mathcal{L}}_c(\{p_{<\mathrm{i}}^{k+1}\}, p_{\mathrm{i}}, \{p_{>\mathrm{i}}^{k}\}, q^k, \{\nu^k\}), \quad (3.11a)$$

$$\nu_{\mathrm{i}}^{k+1} := \nu_{\mathrm{i}}^{k} + c\left(A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^k\right), \quad (3.11b)$$

$$q^{k+1} := \underset{q \in \Omega_q}{\arg\min} \quad \bar{\mathcal{L}}_c(\{p_{\mathrm{i}}^{k+1}\}_{\mathrm{i}=1}^{V}, q, \{\nu^{k+1}\}). \quad (3.11c)$$

Then in the second step we solve (3.10) with two-block ADMM:

$$\boldsymbol{\pi}_{x,\mathrm{i}}^{k+1} := \underset{\boldsymbol{\pi}_{x,\mathrm{i}} \in \Pi_x^{(\mathrm{i})}}{\arg\min} \quad \mathcal{L}_{\mathrm{e},c}(\boldsymbol{\pi}_{x,\mathrm{i}}, \boldsymbol{\pi}_{q,\mathrm{i}}^{k}, \mu_{\mathrm{i}}^{k}), \quad (3.12a)$$

$$\mu_{\mathrm{i}}^{k+1} := \mu_{\mathrm{i}}^{k} + c(A_{\mathrm{e}}^{(\mathrm{i})}\mathrm{vec}(\boldsymbol{\pi}_{x,\mathrm{i}}^{k+1}) - \mathrm{vec}(\boldsymbol{\pi}_{q,\mathrm{i}}^{k})), \quad (3.12b)$$

$$\boldsymbol{\pi}_{y,\mathrm{i}}^{k+1} := \underset{\boldsymbol{\pi}_{y,\mathrm{i}} \in \Pi_y^{(\mathrm{i})}}{\arg\min} \quad \mathcal{L}_{\mathrm{e},c}(\boldsymbol{\pi}_{x,\mathrm{i}}^{k+1}, \boldsymbol{\pi}_{y,\mathrm{i}}, \mu_{\mathrm{i}}^{k+1}), \quad (3.12c)$$

where in (3.11), we define $\{p_{<\mathrm{i}}^{k+1}\} := \{p_l^{k+1}\}_{l=1}^{\mathrm{i}-1}$ to denote the primal variables, up to $\mathrm{i}-1$ views that are already updated to the $(k+1)^{st}$ step, and $\{p_{\mathrm{i}<}^{k}\} := \{p_m^k\}_{m=\mathrm{i}+1}^{V}$ to denote the rest that are still at the k^{th} step. We define $\{p_{<0}^{k+1}\} = \{\varnothing\} = \{p_{>V}^{k}\}$; in (3.11) and (3.12), the superscript k denotes the step index; each of $\Omega_{\mathrm{i}}, \Omega_q, \Pi_x^{(\mathrm{i})}, \Pi_y^{(\mathrm{i})}$ denotes a compound probability simplex. The algorithm starts with (3.11a), updating each view in succession. Then the augmented variables are updated with (3.11c). Finally, the difference between the primal and augmented variables are added to the dual variables (3.11b) to complete the k-th iteration. After (3.11) converges, we perform the complement step (3.12) in similar fashion for each view. This completes the full algorithm.

3.2.2 MvIB: Incremental Update Form

Intuitively, the consensus-complement form works well in the case where the representation overlap, or equivalently the common information, in the observations $\{X\}$ across all views is abundant. However, consider an extreme case where all views are almost distinct, that is, each view is the complement to the others, then the consensus-complement form will be inefficient in forming a consensus. To address this, we propose another formulation of the MvIB that does not form consensus of all views in one instance. By restricting the representation set to $\{Z^{(\mathrm{i})}\}_{\mathrm{i}=1}^{V}$, the proposed incremental-update MvIB Lagrangian is given by:

$$\mathcal{L}_{\text{inc}} := \sum_{\mathrm{i}=1}^{V} \gamma I(\{X\}; Z^{(\mathrm{i})} | \{Z\}_{\mathrm{i}-1}) - I(Y; Z^{(\mathrm{i})} | \{Z\}_{\mathrm{i}-1}). \tag{3.13}$$

Again, to simplify the above, the incremental-update form is subjected to the following constraints:

- For each view $\mathrm{i} \in [V]$, the corresponding representation $Z^{(\mathrm{i})}$ only access $X^{(\mathrm{i})}$, so $Y \to X^{(\mathrm{i})} \to Z^{(\mathrm{i})} \leftarrow \{Z\}_{\mathrm{i}-1}$ forms a Markov chain.

With the restrictions above, in each step, we can replace the observations of all views $\{X\}$ with the view-specific observation $X^{(\mathrm{i})}$ and hence can rewrite (3.13) as:

$$\mathcal{L}_{\text{inc}} := \sum_{\mathrm{i}=1}^{V} \gamma I(X^{(\mathrm{i})}; Z^{(\mathrm{i})} | \{Z\}_{\mathrm{i}-1}) - I(Y; Z^{(\mathrm{i})} | \{Z\}_{\mathrm{i}-1}). \tag{3.14}$$

In solving (3.14), we consider the following algorithm, where at the i^{th} step, we have:

$$P^{(\mathrm{i})}_{z|x,z_{<\mathrm{i}}} := \arg\min_{P\in\Omega^{(\mathrm{i})}} \quad \mathcal{L}_{\mathrm{inc}}(P, \{P^{(\mathrm{j})}_{z|y,z_{<\mathrm{j}}}\}_{\mathrm{j}=1}^{\mathrm{i}-1}), \tag{3.15a}$$

$$p(z^{(\mathrm{i})}|y, z_{<\mathrm{i}}) = \frac{\sum_{x^{(\mathrm{i})}} p(x^{(\mathrm{i})}|y)p(z^{(\mathrm{i})}, z_{<\mathrm{i}}|x^{(\mathrm{i})})}{p(z_{<\mathrm{i}}|y)}, \tag{3.15b}$$

where step (3.15a) is solved with algorithm (3.12) with priors replaced with $P^{(\mathrm{i})}_{z|x,z_{<\mathrm{i}}}$, denoting the tensor form of a conditional probability:

$$p(z^{(\mathrm{i})}|x^{(\mathrm{i})}, z^{(\mathrm{i}-1)}, \cdots, z^{(1)}), \forall \mathrm{i} \in [V].$$

The above shown tensor is the primal variable for step i which belongs to a compound simplex $\Omega^{(\mathrm{i})}$ since it is in fact a conditional probability. In the algorithm, for each step (3.15a), we solve it with the algorithm (3.11) by setting $V = 1$ and treating the estimators from the previous steps as additional priors. This can be easily shown through Bayes' rule and the conditional independence assumption. As an example, consider a two-view case with the conditional probabilities $p(z^{(2)}, x^{(2)}|z^{(1)}) = p(x^{(2)}|z^{(1)})p(z^{(2)}|z^{(1)}, x^{(2)})$, and $p(x^{(2)}|z^{(1)}) = \sum_y p(x^{(2)}|y)p(y|z^{(1)})$.

3.2.3 Multi-Source Privacy Funnel

As the duality of the single model IB and PF, in this part, we present the dual problem of the MvIB, which we call the multi-source privacy funnels (MsPF). This generalization of the single modal PF expands the sensitive information Y into a set of sensitive "files" $\{Y^{(\mathrm{i})}\}_{\mathrm{i}=1}^{S}$ with the joint probability of mapping the private information in each file $Y^{(\mathrm{i})}$ to a public information X, denoted as $p(x, y^{(\mathrm{i})}), \forall \mathrm{i} \in [S]$ available at a server. Then the goal is to find a representation Z that minimizes the privacy leakage of sensitive information in all files $\{Y^{(\mathrm{i})}\}_{\mathrm{i}=1}^{S}$ and maximizes the utility of the common information X accessible to a user. The corresponding Markov chain is $\{Y^{(\mathrm{i})}\} - X - Z$ and the metrics of information leakage

and utility are measured in mutual information $I(Y^{(\mathrm{i})}; Z)$ and $I(X; Z)$ respectively. We can solve the MsPF problem through Lagrangian multiplier methods:

$$\mathcal{L}(p_{z|x}) := I(\{Y\}_{\mathrm{i}=1}^{S}; Z) - \gamma I(Z; X), \tag{3.16}$$

where $\gamma > 0$ the trade-off parameter and the variable to optimize is the conditional probability vector $p(z|x)$. In the following, we assume that the elements in the set ${Y^{(\mathrm{i})}}_{\mathrm{i}=1}^{S}$ are mutually exclusive, that is, $I(Y^{(\mathrm{i})}; X|Y_{\mathrm{i}-1}) = 0$, which corresponds to the case when the sensitive information are non-overlapping, otherwise, one can merge the dependent pairs $Y^{(\mathrm{i})}, Y^{(\mathrm{j})}, \mathrm{i} \neq \mathrm{j} \in [S]$ into a single file. Under this assumption, we have:

$$I(\{Y^{(\mathrm{i})}\}_{\mathrm{i}=1}^{S}; Z) = \sum_{\mathrm{i}=1}^{S} I(Y^{(\mathrm{i})}; Z|Y_{\mathrm{i}-1}), \tag{3.17}$$

where $I(Y^{(\mathrm{i})}; Z|Y_{\mathrm{i}-1}) := I(Y^{(\mathrm{i})}; Z|Y^{(1)}, \cdots, Y^{(\mathrm{i}-1)})$, and $I(Y^{(1)}; Z|Y_0) = I(Y^{(1)}; Z)$ by convention [65]. Substitute (3.17) into (3.16), we obtain the MsPF Lagrangian

$$\begin{aligned} \mathcal{L}_{MsPF} : &= \sum_{\mathrm{i}=1}^{S} I(Y^{(\mathrm{i})}; Z) - \gamma I(X; Z) \\ &= -\sum_{\mathrm{i}=1}^{S} H(Z|Y^{(\mathrm{i})}) + H(Z|X) + (S - \gamma)\, H(Z). \end{aligned} \tag{3.18}$$

Note that since the variable to optimize is $p(z|x) \forall x \in \mathcal{X}, z \in \mathcal{Z}$ and $\{Y^{(\mathrm{i})}\}_{\mathrm{i}=1}^{S} - X - Z$ forms a Markov chain, we have $\forall \mathrm{i} \in [S]$:

$$\begin{aligned} p(z|y^{(\mathrm{i})}) &= \sum_{x} p(z|x) p(x|y^{(\mathrm{i})}), \\ p(z) &= \sum_{x} p(z|x) p(x), \\ p(x) &= \sum_{y^{(\mathrm{i})}} p(x, y^{(\mathrm{i})}). \end{aligned}$$

Again, we propose solving (3.18) with the augmented Lagrangian. By reusing the vector form of the (conditional) probability vectors (3.7), we further define:

$$\begin{aligned} p_{\mathrm{i}} &:= p_{z|y^{(\mathrm{i})}}, \quad q := p_{z|x}, \\ F_{\mathrm{i}}(p_{\mathrm{i}}) &:= -H(Z|Y^{(\mathrm{i})}), \\ G(q) &:= H(Z|X) + (S-\gamma)H(Z), \\ p_{\mathrm{i}} &= B_{\mathrm{i}} q, \quad \forall \mathrm{i} \in [S], \end{aligned} \tag{3.19}$$

where we treat $p(z) := \sum_x p(z|x)p(x)$ as a strict equality constraint and the matrix $B_{\mathrm{i}}, \forall \mathrm{i} \in [S]$ is defined as $B_{\mathrm{i}} := I_Z \otimes P^T_{X|Y^{(\mathrm{i})}}$ with I_Z a $|\mathcal{Z}| \times |\mathcal{Z}|$-dimensional identity matrix, $\otimes$ the Kronecker product, and $P_{X|Y^{(\mathrm{i})}}$ the matrix form of a conditional probability whose (l,m)-th entry is $p(x_l|y^{(\mathrm{i})}_m)$. Hence B_{i} is full row rank and block diagonal. Note that, by construction $F_{\mathrm{i}}(p_{\mathrm{i}})$ is convex whereas $G(q)$ is weakly convex by Lemma 2.5.3. Following the above formulation (3.19), the corresponding augmented Lagrangian is:

$$\mathcal{L}_c(\{p_{\mathrm{i}}\}_{\mathrm{i}=1}^S, q, \{\nu_{\mathrm{i}}\}_{\mathrm{i}=1}^S) := \sum_{\mathrm{i}=1}^S F_{\mathrm{i}}(p_{\mathrm{i}}) + G(q) + \sum_{\mathrm{i}=1}^S \left[\langle \nu_{\mathrm{i}}, p_{\mathrm{i}} - B_{\mathrm{i}} q \rangle + \frac{c}{2} \|p_{\mathrm{i}} - B_{\mathrm{i}} q\|^2 \right], \tag{3.20}$$

where $\{\nu_{\mathrm{i}}\}_{\mathrm{i}=1}^S$ denotes the set of dual variables for all S sources while $c > 0$ is the fixed penalty coefficient. Different from the case in MvIB, our goal for the proposed MsPF is to reduce the complexity at the server to process the sensitive information which scales exponentially as the number of sources S increase with the single modal PF solvers. To avoid the infeasible complexity, we propose solving (3.20) with the following ADMM:

$$p_{\mathrm{i}}^{k+1} := \underset{p_{\mathrm{i}} \in \Omega_p^{(\mathrm{i})}}{\arg\min} \quad \mathcal{L}_c(\{p_{\mathrm{i}}^{k+1}\}_{<\mathrm{i}}, p_{\mathrm{i}}, \{p_{\mathrm{i}}^k\}_{>\mathrm{i}}, q^k, \{\nu_{\mathrm{i}}^k\}_{\mathrm{i}=1}^S), \tag{3.21a}$$

$$\nu_{\mathrm{i}}^{k+1} := \nu_{\mathrm{i}}^k + c\left(p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^k\right), \tag{3.21b}$$

$$q^{k+1} := \underset{q \in \Omega_q}{\arg\min} \quad \mathcal{L}_c(\{p_{\mathrm{i}}\}_{\mathrm{i}=1}^S, q, \{\nu_{\mathrm{i}}^{k+1}\}_{\mathrm{i}=1}^S), \tag{3.21c}$$

where the superscript $k \in \mathbb{N}$ denotes the iteration counter and $\{< \mathrm{i}\}$ denotes the index set $\{\mathrm{j} | \mathrm{j} < \mathrm{i}, \mathrm{j} \in \mathbb{N}\}$ and similarly for $\{> \mathrm{i}\}$ but we define $\{< 1\} = \{\varnothing\} = \{> S\}$. Note that the order in updating the dual variables $\{\nu\}$ is different from that of the MvIB consensus

Table 3.1. Summary of Convergence and Complexity for MvIB and MsPF

Algorithms	Dimension Complexity	Rate of Conv.	Properties of Functions
Consensus Complement	$\mathcal{O}(\|V\|\|X\|\|Z\|^2)$	Locally Linear	F_i: convex, F_i: M_i-Lipschitz Continuous G: σ_G-weakly cvx, L_q-smooth
Incremental Update	$\mathcal{O}(\|X\|\|Z\|^V)$	Locally Linear	F_i: convex F_i: M_i-Lipschitz Continuous G: σ_G-weakly cvx, L_q-smooth
Multi-source Privacy Funnel	$\mathcal{O}(\|S\|\|X\|\|Z\|)$	Locally Linear	F_i: convex, G: σ_G-weakly cvx, L_q-smooth G: M_q-Lipschitz Continuous

algorithm (3.11). Moreover, the difference can be observed from the minimizer conditions of the augmented Lagrangian shown below.

$$\begin{aligned} 0 &= \nabla F_\mathrm{i}(p_\mathrm{i}^{k+1}) + \nu_\mathrm{i}^k + c\left(p_\mathrm{i}^{k+1} - B_\mathrm{i} q^k\right) \\ &= F_\mathrm{i}(p_\mathrm{i}^{k+1}) + \nu_\mathrm{i}^{k+1} \qquad (3.22a) \\ \nu_\mathrm{i}^{k+1} &= \nu_\mathrm{i}^k + c\left(p_\mathrm{i}^{k+1} - B_\mathrm{i} q^k\right), \qquad (3.22b) \\ 0 &= \nabla G(q^{k+1}) - B_\mathrm{i}^T\left[\nu_\mathrm{i}^{k+1} + c\left(p_\mathrm{i}^{k+1} - B_\mathrm{i} q^{k+1}\right)\right]. \qquad (3.22c) \end{aligned}$$

The assumptions and minimizer conditions will be adopted in proving the proposed MSPF algorithm (3.21), which turns out to be Q-linearly fast toward a local stationary point w^*.

3.3 Convergence Analysis

In this part, we prove the convergence of the proposed two algorithms, i.e., (3.13) and (3.11). Observe that in the consensus-complement form, it suffices to prove the convergence of the consensus step (3.11) since the complement step (3.12) and the incremental algorithm (3.15) are special cases with $V = 1$. For convenience of expression, in the rest of the proof, we will denote $\bar{\mathcal{L}}_c^k := \bar{\mathcal{L}}_c(\{p^k\}, q^k, \{\nu^k\})$ the function value evaluated with w^k, where $w^k := (\{p^k\}, q^k, \{\nu^k\})$ the collective point at step k. In consistent to previous chapters, part of the definitions and notations are in consistent with that used in section 2.4.1.

3.3.1 Locally R-Linear Rate of Convergence

In proving the convergence and the locally linear rate of convergence of the proposed MvIB algorithms, we consider the case where the following assumptions are satisfied:

Assumption D.

- *There exists stationary points* $w^* := (\{p_\mathrm{i}^*\}, q^*, \{\nu_\mathrm{i}^*\})$ *that belong to a set* $\Omega^* := \{w | w \in \Omega, \nabla \mathcal{L}_c(w) = 0\}$.
- $F_\mathrm{i}(p_\mathrm{i}), \forall \mathrm{i} \in [V]$ *is* L_i*-smooth,* M_i*-Lipschitz continuous and convex;* $G(q)$ *is* L_q*-smooth and* σ_G*-weakly convex.*
- *The penalty coefficient* c *satisfies:*

$$c > \max_{\mathrm{i} \in [V]} \left\{ \frac{\sigma_G}{V}, \frac{\lambda_{A_\mathrm{i}} L_\mathrm{i} M_\mathrm{i} \sqrt{2}}{\mu_{A_\mathrm{i} A_\mathrm{i}^T}} \right\}.$$

From the first order minimizer conditions, we have:

$$0 = \nabla F_\mathrm{i}(p_\mathrm{i}^{k+1}) + A_\mathrm{i}^T \left[\nu_\mathrm{i}^k + c(A_\mathrm{i} p_\mathrm{i}^{k+1} - q^k) \right], \tag{3.23a}$$

$$\nu_\mathrm{i}^{k+1} = \nu_\mathrm{i}^k + c \left(A_\mathrm{i} p_\mathrm{i}^{k+1} - q^k \right), \tag{3.23b}$$

$$0 = \nabla G(q^{k+1}) - \sum_{\mathrm{j}=1}^{V} \left[\nu_\mathrm{j}^{k+1} + c(A_\mathrm{j} p_\mathrm{j}^{k+1} - q^{k+1}) \right]. \tag{3.23c}$$

Note that at a stationary point w^*, the above-mentioned minimizer conditions (3.23) reduces to the following. $\forall \mathrm{i} \in [V]$:

$$\nabla F_\mathrm{i}(p_\mathrm{i}^*) = -A_\mathrm{i}^T \nu^*, \quad \nabla G(q^*) = V \nu^*, \quad q^* = A_\mathrm{i} p_\mathrm{i}^*. \tag{3.24}$$

The first step of the proof, as in the case for single-modal DRS solvers, is to develop a sufficient decrease lemma.

Lemma 3.3.1. *Suppose Assumption D is satisfied. Define* $\bar{\mathcal{L}}_c(\{p\}, q, \{\nu\})$ *as in* (3.9). *If* (3.9) *is solved with the algorithm* (3.11), *we have:*

$$\bar{\mathcal{L}}_c^k - \bar{\mathcal{L}}_c^{k+1} \geq \sum_{\mathrm{i}=1}^{V} \rho_{p,\mathrm{i}} \|p_\mathrm{i}^k - p_\mathrm{i}^{k+1}\|^2 + \rho_q \|q^k - q^{k+1}\|^2, \tag{3.25}$$

where $\rho_{p,\mathrm{i}} := c/(2M_\mathrm{i}^2) - \lambda_{A_\mathrm{i}}^2 L_\mathrm{i}^2/(c\mu_{A_\mathrm{i}A_\mathrm{i}^T}^2), \rho_q := (cV - \sigma_G)/2$. λ_B, μ_B *denote the largest and smallest singular value of a matrix* B.

Proof. We simply expand the l.h.s. of (3.25) w.r.t. the consecutive update from step k to $k+1$ according to the algorithm (3.11). With a slight abuse of notation, skipping the variables that are fixed in consecutive updates for clarity, by the p_i-update for each view $\mathrm{i} \in [V]$, we have:

$$\begin{aligned}
\bar{\mathcal{L}}_c(p_\mathrm{i}^k) - \bar{\mathcal{L}}_c(p_\mathrm{i}^{k+1}) =& F_\mathrm{i}(p_\mathrm{i}^k) - F_\mathrm{i}(p_\mathrm{i}^{k+1}) + \langle \nu_\mathrm{i}^k, A_\mathrm{i}p^k - A_\mathrm{i}p^{k+1}\rangle \\
&+ \frac{c}{2}\|A_\mathrm{i}p_\mathrm{i}^k - q^k\|^2 - \frac{c}{2}\|A_\mathrm{i}p_\mathrm{i}^{k+1} - q^k\|^2 \\
\geq& \langle \nabla F_\mathrm{i}(p_\mathrm{i}^{k+1}) + A^T\nu_\mathrm{i}^k, p_\mathrm{i}^k - p_\mathrm{i}^{k+1}\rangle + \frac{c}{2}\|A_\mathrm{i}p_\mathrm{i}^k - q^k\|^2 - \frac{c}{2}\|A_\mathrm{i}p_\mathrm{i}^{k+1} - q^k\|^2 \\
=& -c\langle A_\mathrm{i}p_\mathrm{i}^{k+1} - q^k, A_\mathrm{i}p_\mathrm{i}^k - A_\mathrm{i}p_\mathrm{i}^{k+1}\rangle + \frac{c}{2}\|A_\mathrm{i}p_\mathrm{i}^k - q^k\|^2 - \frac{c}{2}\|A_\mathrm{i}p_\mathrm{i}^{k+1} - q^k\|^2 \\
=& \frac{c}{2}\|A_\mathrm{i}p_\mathrm{i}^k - A_\mathrm{i}p_\mathrm{i}^{k+1}\|^2.
\end{aligned} \tag{3.26}$$

where the inequality follows the convexity of F_i and the last equality follows the minimizer condition (3.23a). Then for the q update:

$$\begin{aligned}
\bar{\mathcal{L}}_c(q^k) - \bar{\mathcal{L}}_c(q^{k+1}) =& G(q^k) - G(q^{k+1}) + \sum_{\mathrm{i}=1}^{V} \Big[\langle \nu_\mathrm{i}^{k+1}, q^k - q^{k+1}\rangle \\
&+ \frac{c}{2}\|A_\mathrm{i} p_\mathrm{i}^{k+1} - q^k\|^2 - \frac{c}{2}\|A - \mathrm{i} p_\mathrm{i}^{k+1} - q^{k+1}\|^2\Big] \\
\geq& \langle \nabla G(q^{k+1}) + \sum_{\mathrm{i}=1}^{V} \nu_\mathrm{i}^{k+1}, q^k - q^{k+1}\rangle - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 \\
&+ \frac{c}{2}\|A_\mathrm{i} p_\mathrm{i}^{k+1} - q^k\|^2 - \frac{c}{2}\|A_\mathrm{i} p_\mathrm{i}^{k+1} - q^{k+1}\|^2 \\
=& c\sum_{\mathrm{i}=1}^{V} \langle A_\mathrm{i} p_\mathrm{i}^{k+1} - q^{k+1}, q^k - q^{k+1}\rangle - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 \\
&+ \frac{c}{2}\|A_\mathrm{i} p_\mathrm{i}^{k+1} - q^k\|^2 - \frac{c}{2}\|A_\mathrm{i} p_\mathrm{i}^{k+1} - q^{k+1}\|^2 \\
=& \left(\frac{c - \sigma_G}{2}\right)\|q^k - q^{k+1}\|^2,
\end{aligned} \tag{3.27}$$

where the inequality is due to the weak convexity of G and the last equality follows (3.23c). Lastly, for the dual ascend:

$$\bar{\mathcal{L}}_c(\{\nu^k\}) - \bar{\mathcal{L}}_c(\{\nu^{k+1}\}) = -\frac{1}{c}\sum_{\mathrm{i}=1}^{V}\|\nu_\mathrm{i}^k - \nu_\mathrm{i}^{k+1}\|^2. \tag{3.28}$$

To connect the dual ascend and the p_i-update, consider the following:

$$\begin{aligned}
\|\nu_\mathrm{i}^k - \nu_\mathrm{i}^{k+1}\| &\leq \|(A_\mathrm{i}A_\mathrm{i}^T)^{-1}\|\|A_\mathrm{i}\|\|A_\mathrm{i}^T\nu_\mathrm{i}^k - A_\mathrm{i}^T\nu_\mathrm{i}^{k+1}\| \\
&= \frac{\lambda_{A_\mathrm{i}}}{\mu_{A_\mathrm{i}A_\mathrm{i}^T}}\|\nabla F_\mathrm{i}(p_\mathrm{i}^k) - \nabla F_\mathrm{i}(p_\mathrm{i}^{k+1})\| \leq \frac{\lambda_{A_\mathrm{i}}L_\mathrm{i}}{\mu_{A_\mathrm{i}A_\mathrm{i}^T}}\|p_\mathrm{i}^k - p_\mathrm{i}^{k+1}\|,
\end{aligned} \tag{3.29}$$

where μ_B, λ_B denotes the smallest and largest singular value of a matrix B respectively. (3.29) is because $A_\mathrm{i}, \forall \mathrm{i} \in [V]$ is full row rank, the minimizer condition (3.23a) and F_i being L_i-smooth. In addition, using the sub-minimization path technique [45], observe that the p_i-update is equivalent to a convex proximal operator:

$$W_c(\zeta) := \underset{p_\mathrm{i}\in\Omega_\mathrm{i}}{\arg\min} \quad \mathcal{F}_\mathrm{i}(p_\mathrm{i}) + \frac{c}{2}\|\zeta - A_\mathrm{i}p_\mathrm{i}\|^2, \tag{3.30}$$

where by letting $\zeta = \frac{\nu_{\mathrm{i}}^k}{c} - q^k$, we recover the p_{i}-update (3.11a). Then since p_{i} is within a compact compound simplex and F_{i} is Lipshitz continuous, which in turns means $W(\zeta)$ is M_{i}-Lipschitz continuous:

$$\|W(A_{\mathrm{i}}p^k) - W(A_{\mathrm{i}}p^{k+1})\| = \|p^k - p^{k+1}\| \leq M_{\mathrm{i}}\|A_{\mathrm{i}}p_{\mathrm{i}}^k - A_{\mathrm{i}}p_{\mathrm{i}}^{k+1}\|. \tag{3.31}$$

Applying (3.29) and (3.31) to (3.26), we get:

$$\begin{aligned}
&\bar{\mathcal{L}}_c(\{p_{\mathrm{i}}^k\}, q^k, \{\nu^k\})_{\mathrm{i}=1}^V - \bar{\mathcal{L}}_c(\{p_{\mathrm{i}}^{k+1}\}, q^{k+1}, \{\nu^{k+1}\})_{\mathrm{i}=1}^V \\
\geq &\frac{cV - \sigma_G}{2}\|q^k - q^{k+1}\|^2 + \sum_{\mathrm{i}=1}^V \left[\frac{c}{2}\|A_{\mathrm{i}}p_{\mathrm{i}}^k - A_{\mathrm{i}}p_{\mathrm{i}}^{k+1}\| - \frac{1}{c}\|\nu_{\mathrm{i}}^k - \nu_{\mathrm{i}}^{k+1}\|^2\right] \\
\geq &\frac{cV - \sigma_G}{2}\|q^k - q^{k+1}\|^2 + \sum_{\mathrm{i}=1}^V \left(\frac{c}{2M_{\mathrm{i}}^2} - \frac{\lambda_{A_{\mathrm{i}}}^2 L_{\mathrm{i}}^2}{c\mu_{A_{\mathrm{i}}A_{\mathrm{i}}^T}^2}\right)\|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2.
\end{aligned}$$

Finally, to make the coefficients pre-multiplied to the squared norm positive:

$$c > \max_{\mathrm{i}\in[V]}\left\{\frac{\sigma_G}{V}, \frac{\lambda_{A_{\mathrm{i}}}L_{\mathrm{i}}M_{\mathrm{i}}\sqrt{2}}{\mu_{A_{\mathrm{i}}A_{\mathrm{i}}^T}}\right\}.$$

□

Note that Lemma 3.3.1 applies to any objective function that can be decomposed into a convex-weakly convex pair including a special case in our formulation with $V = 1$. Remarkably, $V = 1$ also applies to the complement step (3.12) and each individual step for all V views of the incremental algorithm (3.15). The sufficient decrease lemma implies convergence of the sequence $\{w^k\}_{k\in\mathcal{N}}$ obtained through the proposed algorithms. Moreover, it also implies that the sequence of the function values $\{\bar{\mathcal{L}}_c^k\}$ converges R-linearly. To prove this, we find an upper bound of the function value difference between the evaluation of w^* and that of the feasible solutions around its neighborhood, which gives the next lemma.

Lemma 3.3.2. *Define* $w^k := (\{A_i p_i^k\}_{i=1}^V, q^k, \{\nu_i^k\}_{i=1}^V)$ *the collective point at step* k *and let* w^* *be a stationary point. We have:*

$$\bar{\mathcal{L}}_c^{k+1} - \bar{\mathcal{L}}_c^* \leq \frac{c}{2}\|q^{k+1} - q^k\|^2 + \frac{\sigma_G}{2}\|q^{k+1} - q^*\|^2 + \frac{1}{c}\sum_{i=1}^V \|\nu_i^{k+1} - \nu_i^k\|^2.$$

Proof. Consider the following:

$$\begin{aligned}\bar{\mathcal{L}}_c^{k+1} - \bar{\mathcal{L}}_c^* =& \sum_{i=1}^V F_i(p_i^{k+1}) - F_i(p_i^*) + \langle \nu_i^{k+1}, A_i p_i^{k+1} - q^{k+1}\rangle \\ &- \langle \nu_i, A_i p_i^* - q^*\rangle + G(q^{k+1}) - G(q^*).\end{aligned} \tag{3.32}$$

Then we separately consider each term $F_i, \forall i \in [V]$:

$$\begin{aligned}F_i(p_i^{k+1}) - F_i(p_i^*) \leq& \langle \nabla F_i(p_i^{k+1}), p_i^{k+1} - p_i^*\rangle \\ =& -\langle \nu_i^k + c\left(A_i p_i^{k+1} - q^k\right), A_i p_i^{k+1} - A_i p_i^*\rangle,\end{aligned}$$

where the inequality is due to convexity of F_i while the equality is by the minimizer condition (3.23a). Similarly, by the σ_G-weak convexity of G, we have:

$$G(q^{k+1}) - G(q^*) \leq \sum_{i=1}^V \langle \nu_i^{k+1}, q^{k+1} - q^*\rangle + \frac{\sigma_G}{2}\|q^{k+1} - q^*\|^2.$$

Substitute the above into (3.32), then using the identity (2.14) and the stationary point conditions (3.24), we get:

$$\bar{\mathcal{L}}_c^{k+1} - \bar{\mathcal{L}}_c^* \leq \frac{c}{2}\|q^{k+1} - q^k\|^2 + \frac{\sigma_G}{2}\|q^{k+1} - q^*\|^2 + c\sum_{i=1}^V \|A_i p_i^{k+1} - q^{k+1}\|^2,$$

where the last term in the above follows from (3.11b), that is $\nu_i^{k+1} - \nu_i^k = c(A_i p_i^{k+1} - q^{k+1})$, and we complete the proof. □

Combining Lemma 3.3.2 with the sufficient decrease lemma, we can derive the local linear rate of convergence of the consensus step (3.11) of the proposed MvIB algorithm, and since the incremental-update algorithm is a special case with $V = 1$, the proven results apply the it as well.

Theorem 3.3.3. *Suppose Assumption D is satisfied. Define $\bar{\mathcal{L}}_c$ as in (3.9) and solved with the algorithm (3.11). If the penalty coefficient $c > c_{\min}$, then w^k converges to a stationary point w^* linearly around a neighborhood of w^* such that $\|w - w^*\|^2 < \varepsilon$, $\bar{\mathcal{L}}_c^* < \bar{\mathcal{L}}_c < \bar{\mathcal{L}}_c^* + \delta$, where $\varepsilon, \delta > 0$.*

Proof. Define $w^k := (\{A_{\mathrm{i}} p_{\mathrm{i}}^k\}_{\mathrm{i}=1}^V, q^k, \{\nu_{\mathrm{i}}^k\}_{\mathrm{i}=1}^V)$ the collective point at step k. By Lemma 3.3.1, if the penalty coefficient c satisfies D, then there exists a constant $\tau_1 > 0$ such that:

$$\bar{\mathcal{L}}_c^k - \bar{\mathcal{L}}_c^{k+1} \geq \tau_1 \left[\|q^k - q^{k+1}\|^2 + \sum_{\mathrm{i}=1}^V \|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2 \right]. \tag{3.33}$$

On the other hand, by Lemma 3.3.2, there exists a $\tau_2 > 0$ such that:

$$\begin{aligned} \bar{\mathcal{L}}_c^{k+1} - \bar{\mathcal{L}}_c^* \leq & \tau_2 \left[\|q^{k+1} - q^*\|^2 + \|q^{k+1} - q^k\|^2 + \sum_{\mathrm{i}=1}^V \|\nu_{\mathrm{i}}^{k+1} - \nu_{\mathrm{i}}^k\|^2 \right] \\ \leq & \tau_2 \left[\|q^{k+1} - q^*\|^2 + \|q^{k+1} - q^k\|^2 + \sum_{\mathrm{i}=1}^V \frac{L_{\mathrm{i}}^2}{\mu_{A_{\mathrm{i}} A_{\mathrm{i}}^T}} \|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2 \right] \\ \leq & \tau_3 \left[\|q^{k+1} - q^*\|^2 + \|q^{k+1} - q^k\|^2 + \sum_{\mathrm{i}=1}^V \|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2 \right], \end{aligned} \tag{3.34}$$

where the second inequality is due to (3.29) and we define the constant:

$$\tau_3 := \tau_2 \max_{\mathrm{i} \in [V]} \{1, L_{\mathrm{i}}^2 / \mu_{A_{\mathrm{i}} A_{\mathrm{i}}^T}\} > 0.$$

Substitute (3.33) into (3.34), we have:

$$\bar{\mathcal{L}}_c^{k+1} - \bar{\mathcal{L}}_c^* \leq \tau_3 \left(\bar{\mathcal{L}}_c^k - \bar{\mathcal{L}}_c^* - \left(\bar{\mathcal{L}}_c^{k+1} - \bar{\mathcal{L}}_c^* \right) + \|q^{k+1} - q^*\|^2 \right).$$

Rearranging the above, we get:

$$\frac{\bar{\mathcal{L}}_c^{k+1} - \bar{\mathcal{L}}_c^*}{\bar{\mathcal{L}}_c^k - \bar{\mathcal{L}}_c^*} \leq \left(\frac{\tau_3}{1 + \tau_3} \right) + \left[\left(\frac{\tau_3}{1 + \tau_3} \right) \frac{\|q^{k+1} - q^*\|^2}{\bar{\mathcal{L}}_c^k - \bar{\mathcal{L}}_c^*} \right].$$

From Lemma 3.3.1, for c sufficient large, $\bar{\mathcal{L}}_c^k - \bar{\mathcal{L}}_c^* \geq \bar{\mathcal{L}}_c^{k+1} - \bar{\mathcal{L}}_c^*$ holds. In addition, by the definition of a neighborhood around w^*, there exist $\varepsilon, \delta, \xi > 0$ such that

$$\delta > \bar{\mathcal{L}}_c^{k+1} - \bar{\mathcal{L}}_c^* > \xi > 0,$$

for $\|w^{k+1} - w^*\| < \varepsilon, \bar{\mathcal{L}}_c^* < \bar{\mathcal{L}}_c^{k+1} < \bar{\mathcal{L}}_c^* + \delta$.

By choosing ε, δ, such that $\varepsilon < \sqrt{\xi/\tau_3} < \sqrt{\delta/\tau_3}$:

$$\frac{\|q^{k+1} - q^*\|^2}{\bar{\mathcal{L}}_c^k - \bar{\mathcal{L}}_c^*} \leq \frac{\varepsilon^2}{\xi} < \frac{1}{\tau_3}$$

holds for k sufficiently large, say $k > N_0, k, N_0 \in \mathbb{N}$. Then the sequence $\{\bar{\mathcal{L}}_c^{\mathrm{i}}\}_{\mathrm{i} \geq N_0}$ is Q-linearly convergent for each view $\mathrm{i} \in [V]$. In turns, this implies that $\sum_{k=1}^{\infty} \|w^{k+1} - w^k\|^2$ is finite, and hence there exist $\{M_p^{(\mathrm{i})}\}_{\mathrm{i}=1}^V, M_q, \{M_\nu^{(\mathrm{i})}\} > 0$, where $0 < Q < 1$, such that for $k > N_0$ and $\forall \mathrm{i} \in [V]$:

$$\begin{aligned}
\rho_p \|A_{\mathrm{i}} p_{\mathrm{i}}^k - A_{\mathrm{i}} p_{\mathrm{i}}^{k+1}\|^2 &\leq \bar{\mathcal{L}}_c^k - \bar{\mathcal{L}}^* \leq M_p^{(\mathrm{i})} Q^k, \\
\rho_q \|q^k - q^{k+1}\|^2 &\leq \bar{\mathcal{L}}_c^k - \bar{\mathcal{L}}^* \leq M_q Q^k, \\
\rho_\nu \|\nu_{\mathrm{i}}^k - \nu_{\mathrm{i}}^{k+1}\|^2 &\leq \bar{\mathcal{L}}_c^k - \bar{\mathcal{L}}^* \leq M_\nu^{(\mathrm{i})} Q^k,
\end{aligned}$$

where the last inequality follows (3.23c) and the L_{i} smoothness of each F_{i} and (3.29). Combining the above we get:

$$\|w^m - w^n\|^2 \leq \sum_{k=m}^{n-1} \|w^k - w^{k+1}\| \leq \frac{\bar{M} Q^n}{\bar{\rho}(1-Q)},$$

where $\bar{M} := M_q + \sum_{\mathrm{i}=1}^V M_p^{(\mathrm{i})} + M_\nu^{(\mathrm{i})}$, $\bar{\rho} := \min\{\rho_p, \rho_q, \rho_\nu\}$. Then since the above is a Cauchy sequence, by taking limit with $m \to \infty$, that is, $w^m \to w^*$ we have:

$$\|w^n - w^*\|^2 \leq \frac{\bar{K} Q^n}{\bar{\rho}(1-Q)},$$

which proves that $\{w^n\}_{n > N_0}$ is R-linearly convergent. □

As a remark, the convergence does not require any strong convexity of $\{F_{\mathrm{i}}\}_{\mathrm{i}=1}^{V}$ but convexity is still required. As for the convergence rate, Theorem 3.3.3 implies that the convergence is not point wise, as each of the components $\{p_{\mathrm{i}}^k\}_{\mathrm{i}=1}^{V}$ in the collection point w^k is pre-multiplied by the matrices $A_{\mathrm{i}}, \forall \mathrm{i} \in [V]$. In turns, this hints that it is possible to have multiple solutions that give the same MvIB losses. To interpret this, recall that the primal variables $\{p_{\mathrm{i}}\}_{\mathrm{i}=1}^{V}$ and the augmented Lagrangian is consists of a combination of mutual information. As mutual information is invariant to symmetry, point-wise convergence is not necessary. Finally, in this part, while the convergence analysis is focused on MvIB, but following similar steps in the proof, it can be shown that the MsPF algorithm can also achieve R-linear rate of convergence. However, as will be shown in the next section, it turns out that we can adopt KŁ inequality of prove Q-linear rates of convergence for both MvIB and MsPF. As Q-linear rate is a stronger sense of rate of convergence than R-linear, we therefore defer the convergence analysis of MsPF to the next section.

3.3.2 Locally Q-Linear Rate of Convergence through the KŁ Inequality

As in Chapter 2.4.3, beyond R-linear rate of convergence, we can further improve the theoretic rate of convergence guarantee to Q-linear through the KŁ inequality. Note that as in the remark of the last chapter, both the R-linear and Q-linear rates of convergence results are needed if we generalize the class of consensus algorithms (3.11) to DRS splitting methods (with the relaxation $0 < \alpha < 2$). However, instead of complicating the discussion, we will focus on ADMM ($\alpha = 1$) for illustration purposes, the more general case that corresponds to the DRS methods can be obtained through our results in single modal settings as presented in the last chapter.

To apply the KŁ inequality, we need the additional results, presented the in following, beyond the sufficient decrease lemma (Lemma 3.3.1). Similar to the single modal case, the goal is to show that the Łojasiewicz exponent is 1/2 locally around a stationary point $w^* := (\{p_{\mathrm{i}}^*\}, q^*, \{\nu_{\mathrm{i}}^*\})_{\mathrm{i}=1}^{V}$. The first step is to find an upper bound of the difference of the augmented Lagrangian (3.9).

Lemma 3.3.4. *Let $\bar{\mathcal{L}}_c^k$ be defined as in (3.9). Denote $w^k := (\{p_{\mathrm{i}}^k\}, q^k, \{\nu_{\mathrm{i}}^k\})_{\mathrm{i}=1}^V$ the k-th entry of a sequence $\{w^k\}_{k\in\mathbb{N}}$ obtained through the algorithm (3.11) and denote w^* a stationary point of $\bar{\mathcal{L}}_c$ whose local neighbor is such that $\bar{\mathcal{L}}_c^* < \bar{\mathcal{L}}_c < \bar{\mathcal{L}}_c^* + \delta$ for all $\|w - w^*\| < \varepsilon$ with some constants $\delta, \varepsilon > 0$. Then we have:*

$$\bar{\mathcal{L}}_c^{k+1} - \bar{\mathcal{L}}_c^* \leq \frac{c}{2}\sum_{\mathrm{i}=1}^{V}\|A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - A_{\mathrm{i}}p_{\mathrm{i}}^*\|^2.$$

Proof. For simplicity of expression, in the following, we denote $\mathcal{L}_c^k : \bar{\mathcal{L}}_c(w^k)$. By the augmented Lagrangian (3.9), we have:

$$\begin{aligned}\mathcal{L}_c^{k+1} - \mathcal{L}_c = G(q^{k+1}) + \sum_{\mathrm{i}=1}^{V} F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}) + \langle \nu_{\mathrm{i}}^{k+1}, A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^{k+1}\rangle + \frac{c}{2}\|A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^{k+1}\|^2 \\ - \left[G(q) + \sum_{\mathrm{i}=1}^{V} F_{\mathrm{i}}(p_{\mathrm{i}}) + \langle \nu_{\mathrm{i}}, A_{\mathrm{i}}p_{\mathrm{i}} - q\rangle + \frac{c}{2}\|A_{\mathrm{i}}p_{\mathrm{i}} - q\|^2\right] \quad (3.35)\end{aligned}$$

First, for the $F_{\mathrm{i}}(p_{\mathrm{i}}), \forall \mathrm{i} \in [V]$, by the convexity of F_{i} and the minimizer conditions (3.23a), we have:

$$\begin{aligned}F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}) - F_{\mathrm{i}}(p_{\mathrm{i}}) &\leq \langle \nabla F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}), p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}}\rangle \\ &= \langle -A_{\mathrm{i}}^T\left[\nu_{\mathrm{i}}^k + c\left(A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^k\right)\right], p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}}\rangle \quad (3.36)\\ &= c\langle \nu_{\mathrm{i}}^{k+1}, A_{\mathrm{i}}p_{\mathrm{i}} - A_{\mathrm{i}}p_{\mathrm{i}}^{k+1}\rangle,\end{aligned}$$

where the first inequality follows the convexity of $F_{\mathrm{i}}, \forall \mathrm{i} \in [V]$. Similarly, for the other sub-objective function $G(q)$, by its weak convexity:

$$\begin{aligned}G(q^{k+1}) - G(q) &\leq \langle \nabla G(q^{k+1}), q^{k+1} - q\rangle + \frac{\sigma_G}{2}\|q^{k+1} - q\|^2 \\ &= \sum_{\mathrm{i}=1}^{V}\langle \nu_{\mathrm{i}}^{k+1} + c\left(A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^{k+1}\right), q^{k+1} - q\rangle + \frac{\sigma_G}{2}\|q^{k+1} - q\|^2,\end{aligned} \quad (3.37)$$

where the first inequality is due to the σ_G-weak convexity of $G(q)$ and we last equality follows the minimizer condition (3.23c). Substitute (3.36) and (3.37) into (3.35), we have:

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c \leq \sum_{\mathrm{i}=1}^{V} \langle \nu_{\mathrm{i}}^{k+1}, A_{\mathrm{i}} p_{\mathrm{i}} - q \rangle + c \sum_{\mathrm{i}=1}^{V} \langle A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - q^{k+1}, q^{k+1} - q \rangle + \frac{\sigma_G}{2} \|q^{k+1} - q\|^2 \\ + \frac{c}{2} \sum_{\mathrm{i}=1}^{V} \left[\|A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - q^{k+1}\|^2 - \|A_{\mathrm{i}} p_{\mathrm{i}} - q\|^2 \right].$$

Substitute w with w^*, and note that at w^*, the relation $A_{\mathrm{i}} p_{\mathrm{i}}^* = q^*, \forall \mathrm{i} \in [V]$. Putting these together, we have:

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq c \sum_{\mathrm{i}=1}^{V} \langle A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - q^{k+1}, q^{k+1} - q^* \rangle + \frac{\sigma_G}{2} \|q^{k+1} - q^*\|^2 + \frac{c}{2} \sum_{\mathrm{i}=1}^{V} \|A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - q^{k+1}\|^2. \tag{3.38}$$

Then for each $\mathrm{i} \in [V]$, we expand the inner product by the identity (2.14) as follows:

$$2 \sum_{\mathrm{i}=1}^{V} \langle A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - q^{k+1}, q^{k+1} - q^* \rangle = \sum_{\mathrm{i}=1}^{V} \left[\|A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - A_{\mathrm{i}} p_{\mathrm{i}}^*\|^2 - \|A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - q^{k+1}\|^2 - \|q^{k+1} - q^*\|^2 \right].$$

Substitute the above into (3.38), we get:

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \frac{c}{2} \sum_{\mathrm{i}=1}^{V} \|A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - A_{\mathrm{i}} p_{\mathrm{i}}^*\|^2 + \frac{\sigma_G - cV}{2} \|q^{k+1} - q^*\|^2 \leq \frac{c}{2} \sum_{\mathrm{i}=1}^{V} \|A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - A_{\mathrm{i}} p_{\mathrm{i}}^*\|^2,$$

where the last inequality is by the range of c in Assumption D. □

Then the second step is to establish the lower bound of the gradient norm of the augmented Lagrangian, which is accomplished by the next lemma.

Lemma 3.3.5. *Let the augmented Lagrangian $\bar{\mathcal{L}}_c$ be defined as in (3.9), denote $\{w^k\}_{k \in \mathcal{N}}$ the sequence obtained from the algorithm (3.11) with $w^k := (\{p_{\mathrm{i}}^k\}, q^k, \{\nu_{\mathrm{i}}^k\})_{\mathrm{i}=1}^V$ denotes the k-th entry of the sequence. Then we have:*

$$\|\nabla \bar{\mathcal{L}}_c^{k+1}\|^2 \geq \sum_{\mathrm{i}=1}^{V} \eta_{\mathrm{i}} \|A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - q^{k+1}\|^2,$$

where $\eta_{\mathrm{i}} := \mu^*_{A_{\mathrm{i}}A_{\mathrm{i}}^T}(c^2+1), \forall \mathrm{i} \in [V]$*, with the constants defined as* $\mu^*_{AA^T} := \min_{v\in[V]} \mu_{A_vA_v^T}$ *and* μ_B *denotes the smallest singular value of a matrix* B*.*

Proof. For convenience of expression, in the following, we denote $\mathcal{L}_c^k := \mathcal{L}_c(w^k)$ the augmented Lagrangian evaluated at the k-th step collective point $w^k := (\{p_{\mathrm{i}}^k\}, q^k, \{\nu_{\mathrm{i}}^k\})_{\mathrm{i}=1}^V$. With the notation and by the definition of the augmented Lagrangian (3.9), we have:

$$\begin{aligned}\nabla\mathcal{L}_c^{k+1} &= \begin{bmatrix} \nabla F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}) + A_{\mathrm{i}}^T\left[\nu_{\mathrm{i}}^{k+1} + c\left(A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^{k+1}\right)\right] \\ \nabla G(q^{k+1}) - \sum_{\mathrm{i}=1}^V\left[\nu_{\mathrm{i}}^{k+1} + c\left(A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^{k+1}\right)\right] \\ A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^{k+1} \end{bmatrix} \\ &= \begin{bmatrix} cA_{\mathrm{i}}^T\left(A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^{k+1}\right) \\ 0 \\ A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^{k+1} \end{bmatrix},\end{aligned} \tag{3.39}$$

where we substitute the minimizer conditions (3.23) into (3.39). Then since A_{i} is full row rank, there exists a positive singular value $\mu_{A_{\mathrm{i}}A_{\mathrm{i}}^T} > 0$, and hence we have:

$$\begin{aligned}\|\nabla\mathcal{L}_c^{k+1}\|^2 &\geq \sum_{\mathrm{i}=1}^V (c^2+1)\|A_{\mathrm{i}}^T\left(A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^{k+1}\right)\|^2 \\ &\geq \sum_{\mathrm{i}=1}^V \mu_{A_{\mathrm{i}}A_{\mathrm{i}}^T}(c^2+1)\|A_{\mathrm{i}}p_{\mathrm{i}}^{k+1} - q^{k+1}\|^2.\end{aligned}$$

By defining $\eta_{\mathrm{i}} := \mu_{A_{\mathrm{i}}A_{\mathrm{i}}^T}(c^2+1)$, we complete the proof. □

By combining Lemma 3.3.4 and Lemma 3.3.5, around a local neighborhood of a stationary point w^*, we obtain the desired results, that is, proving the Łojasiewicz exponent θ is 1/2. The exponent $\theta = 1/2$, in turns, allows us to adopt the KŁ to prove locally Q-linear rate of convergence [36], [37]. We present the next lemma to show that the Łojasiewicz exponent $\theta = 1/2$ for the augmented Lagrangian defined in (3.9), solved with the proposed consensus-complement algorithm (3.11).

Lemma 3.3.6. *Let* $\bar{\mathcal{L}}_c$ *be defined as in* (3.9)*. Suppose Assumption D is satisfied and the sequence* $\{w^k\}_{k\in\mathbb{N}}$ *is obtained through the algorithm* (3.11) *where* $w^k := (\{p_{\mathrm{i}}^k\}, q^k, \{\nu_{\mathrm{i}}^k\})_{\mathrm{i}=1}^V$

denotes the collective point at step k. Then the Łojasiewicz exponent of the augmented Lagrangian $\theta = 1/2$ locally around a neighborhood of a stationary point w^* with $\bar{\mathcal{L}}_c^* < \bar{\mathcal{L}}_c^{k+1} < \bar{\mathcal{L}}_c^* + \delta$ and $\|w^{k+1} - w^*\| < \varepsilon$ for some constants $\delta, \varepsilon > 0$.

Proof. Again, for the convenience of expression, we denote $\mathcal{L}_c^k := \bar{\mathcal{L}}_c(w^k)$. Then, by Lemma 3.3.4 and assumption D there must exists a constant $S_{\max} > 0$ and a stationary point w^* such that:

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq S_{\max} \left[\sum_{\mathrm{i}=1}^{V} \|A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - A_{\mathrm{i}} p_{\mathrm{i}}^*\|^2 + \|A_{\mathrm{i}} p_{\mathrm{i}}^{k+1} - q^{k+1}\|^2 \right]. \tag{3.40}$$

On the other hand, by Lemma 3.3.5, there also exists another constant $S_{\min} = \min_{\mathrm{i} \in [V]} \{\eta_{\mathrm{i}}\} > 0$ such that:

$$\|\nabla \mathcal{L}_c^{k+1}\|^2 \geq S_{\min} \left[\sum_{\mathrm{i}=1}^{V} \|A_{\mathrm{i}} p^{k+1} - q^{k+1}\|^2 \right]. \tag{3.41}$$

Combining (3.40) and (3.41), we have:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* &\leq \frac{S_{\max}}{S_{\min}} \|\nabla \mathcal{L}_c^{k+1}\|^2 + S_{\max} \|q^{k+1} - q^*\|^2 \\
&= \frac{S_{\max}}{S_{\min}} \|\nabla \mathcal{L}_c^{k+1}\|^2 \left(1 + \frac{S_{\min} \|q^{k+1} - q^k\|^2}{\|\nabla \mathcal{L}_c^{k+1}\|^2} \right) \\
&\leq \frac{S_{\max}}{S_{\min}} \|\nabla \mathcal{L}_c^{k+1}\|^2 \left(1 + \frac{S_{\min} \|w^{k+1} - w^k\|^2}{\|\nabla \mathcal{L}_c^{k+1}\|^2} \right),
\end{aligned}$$

where in the last inequality, we follow the definition of a collective point w^k at step k. Then we apply Lemma 2.4.11, which implies $\eta > \|\nabla \mathcal{L}_c^{k+1}\| > \xi > 0$ for some small constants $\eta, \xi > 0$ around the neighborhood of w^*, as defined in the statement. Putting these together, we have:

$$\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \frac{S_{\max}}{S_{\min}} \|\mathcal{L}_c^{k+1}\|^2 \left(1 + \frac{S_{\min} \|w^{k+1} - w^*\|^2}{\|\nabla \mathcal{L}_c^{k+1}\|^2} \right) \leq \frac{S_{\max}}{S_{\min}} \|\mathcal{L}_c^{k+1}\|^2 \left(1 + \frac{S_{\min} \varepsilon^2}{\xi^2} \right).$$

Define the constant $S^* := (S_{\max}/S_{\min})(1 + S_{\min}\varepsilon^2/\xi^2) > 0$ and taking square-root on both sizes of the above inequality which implies $\theta = 1/2$ by Definition (2.4.6). □

Given the lemmas proved in this part, along with the sufficient decrease lemma, we can then apply the KŁ inequality to show locally linear rate.

Theorem 3.3.7. *Let $\bar{\mathcal{L}}_c$ defined as in (3.9). Suppose Assumption D is satisfied and the sequence $\{w_k\}_{k\in\mathbb{N}}$ is obtained through the algorithm (3.11) where $w^k := (\{p_{\mathrm{i}}^k\}, q^k, \{\nu_{\mathrm{i}}^k\})$ denotes the collective point at step k. Then the sequence $\{w^k\}_{k\in\mathbb{N}}$ converges Q-linearly toward a local stationary point w^* around its neighborhood with $\bar{\mathcal{L}}_c^* < \bar{\mathcal{L}}_c < \bar{\mathcal{L}}_c^* + \delta$ and $\|w - w^*\| < \varepsilon$ for some constants $\delta, \varepsilon > 0$.*

Proof. By Lemma 3.3.6, the Łojasiweicz exponent $\theta = 1/2$ for the augmented Lagrangian (3.9) solved with the algorithm (3.11). As in the single modal case, we need to find the relation: $\|\nabla\bar{\mathcal{L}}_c^k\| \leq S^*\|w^k - w^{k-1}\|, S^* > 0$, and then by the KŁ-inequality (Lemma 2.4.10), the sequence $\{w^k\}_{k\in\mathbb{N}}$ converges to a stationary point w^* locally at Q-linear rate, which is the desired result.

First, by (3.39), we have:

$$\|\nabla\bar{\mathcal{L}}(w^k)\|^2 = \left(c^2+1\right)\sum_{\mathrm{i}=1}^{V}\|A_{\mathrm{i}}^T\left[A_{\mathrm{i}}p_{\mathrm{i}}^k - q^k\right]\| \leq \left(c^2+1\right)\sum_{\mathrm{i}=1}^{V}\lambda_{A_{\mathrm{i}}A_{\mathrm{i}}^T}\|A_{\mathrm{i}}p_{\mathrm{i}}^k - q^k\|^2. \tag{3.42}$$

where λ_A denotes the largest eigenvalue of a matrix A. For the first terms in the above inequality, by Cauchy-Schwarz inequality:

$$\|A_{\mathrm{i}}p_{\mathrm{i}}^k - q^{k-1} + q^{k-1} - q^k\|^2 \leq 2\|A_{\mathrm{i}}p_{\mathrm{i}}^k - q^{k-1}\|^2 + 2\|q^k - q^{k-1}\|^2. \tag{3.43}$$

On the other hand, for the second term, by the minimizer condition (3.23a) and the L_{i}-smoothness of each sub-objective function $F_{\mathrm{i}}(p_{\mathrm{i}})$ along with the condition that A_{i} is full row rank, we have:

$$\mu_{A_{\mathrm{i}}A_{\mathrm{i}}^T}\|\nu_{\mathrm{i}}^k - \nu_{\mathrm{i}}^{k-1}\|^2 \leq \|A_{\mathrm{i}}^T\nu_{\mathrm{i}}^k - A_{\mathrm{i}}^T\nu_{\mathrm{i}}^{k-1}\|^2 \leq L_{\mathrm{i}}^2\|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k-1}\|^2, \tag{3.44}$$

where the second inequality is by the connection $\nabla F(p_{\mathrm{i}}^k) = A_{\mathrm{i}}^T \nu_{\mathrm{i}}^k$. Substitute (3.40) and (3.41) into (3.42), we get:

$$\begin{aligned}\|\nabla \bar{\mathcal{L}}_c(w^k)\|^2 &\leq 2(c^2+1)\sum_{\mathrm{i}=1}^{V} \frac{\lambda_{A_{\mathrm{i}}A_{\mathrm{i}}^T} L_{\mathrm{i}}^2}{\mu_{A_{\mathrm{i}}A_{\mathrm{i}}^T}} \|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k-1}\|^2 + 2(c^2+1)\lambda_{A_{\mathrm{i}}A_{\mathrm{i}}^T}\|q^k - q^{k-1}\|^2 \\ &\leq S_c \left(\sum_{\mathrm{i}=1}^{V} \|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k-1}\|^2 + \|q^k - q^{k-1}\|^2\right) \\ &\leq S_c \|w^k - w^{k-1}\|^2.\end{aligned}$$

where the positive constant S_c is defined as:

$$S_c := 2(c^2+1)\max_{\mathrm{i}\in[V]} \left\{ \frac{\lambda_{A_{\mathrm{i}}A_{\mathrm{i}}^T} L_{\mathrm{i}}^2}{\mu_{A_{\mathrm{i}}A_{\mathrm{i}}^T}}, \lambda_{A_{\mathrm{i}}A_{\mathrm{i}}^T} \right\}.$$

Then, by taking square root on the both sides of the inequality above , we obtain the desired relation $\|\bar{\mathcal{L}}_c(w^k)\| \leq S_c\|w^k - w^{k-1}\|$. Then we can apply Lemma 2.4.10 which completes the proof. □

Theorem 3.3.7 shows that the rates of convergence for the proposed MvIB algorithms (3.11)(3.15) are Q-linear. However, the results cannot be directly generalized to the proposed MsPF algorithm (3.21). This is due to limitations of the linear constraints, $p_{\mathrm{i}} - B_{\mathrm{i}}q, \forall \mathrm{i} \in [S]$, as B_{i} is now full-row rank while in MvIB it is the identity matrix. Our next goal is to prove the Q-linear rate of convergence of the MsPF algorithm. As in previous cases, we start with developing the sufficient decrease lemma based on the following set of assumptions.

Assumption E.

- *There exists stationary points* $w^* := (\{p_{\mathrm{i}}^*\}, q^*, \{\nu_{\mathrm{i}}^*\})_{\mathrm{i}=1}^S$ *that belong to a set* $\Omega^* := \{w | w \in \Omega, \nabla \mathcal{L}_c(w) = 0\}$ *where* $\mathcal{L}_c$ *is defined in* (3.20).
- $F_{\mathrm{i}}(p_{\mathrm{i}})$ *is* L_{i}*-smooth* $\forall \mathrm{i} \in [S]$ *and convex while* $G(q)$ *is* L_q*-smooth,* σ_G*-weakly convex and* M_q*-Lipschitz continuous.*
- *The fixed penalty coefficient satisfies* $c > \max\{\sqrt{2}L_{\mathrm{i}}, M_q^2\sigma_G\}, \forall \mathrm{i} \in [S]$.

Lemma 3.3.8. *Let $\mathcal{L}_c$ be defined as in (3.20). Suppose Assumption E is satisfied. Define $w^k := (\{p_{\mathrm{i}}^k\}, q^k, \{\nu_{\mathrm{i}}^k\})_{\mathrm{i}=1}^S$ the collective point at step k obtained through the algorithm 3.21. Then we have:*

$$\mathcal{L}_c(w^k) - \mathcal{L}_c(w^{k+1}) \geq \sum_{\mathrm{i}=1}^{S} \left(\frac{c}{2} - \frac{L_{\mathrm{i}}^2}{c} \right) \|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2 + \left(\frac{c}{2M_q^2} - \frac{\sigma_G}{2} \right) \|q^k - q^{k+1}\|^2.$$

Proof. For convenience of expression, we will denote $\mathcal{L}_c^k := \mathcal{L}_c(w^k)$ where w^k is the collective point defined in the statement. We separate the difference according to the steps in the ADMM algorithm (3.21) as follows. First, for the convex $F_{\mathrm{i}}, \forall \mathrm{i} \in [S]$:

$$\begin{aligned}
&\mathcal{L}_c(\{p_{\mathrm{i}}^{k+1}\}_{<\mathrm{i}}, p_{\mathrm{i}}^k, \{p_{\mathrm{i}}^k\}_{>\mathrm{i}}, q^k, \nu_{\mathrm{i}}^k) - \mathcal{L}_c(\{p_{\mathrm{i}}^{k+1}\}_{<\mathrm{i}}, p_{\mathrm{i}}^{k+1}, \{p_{\mathrm{i}}^k\}_{>\mathrm{i}}, q^k, \nu_{\mathrm{i}}^k) \\
&\qquad = F_{\mathrm{i}}(p_{\mathrm{i}}^k) - F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}) + \langle \nu_{\mathrm{i}}^k, p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1} \rangle + \frac{c}{2} \|p_{\mathrm{i}}^k - B_{\mathrm{i}} q^k\|^2 - \frac{c}{2} \|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^k\|^2.
\end{aligned}$$

Then by the convexity of $F_{\mathrm{i}}, \forall \mathrm{i} \in [S]$:

$$\begin{aligned}
F_{\mathrm{i}}(p_{\mathrm{i}}^k) - F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}) + \langle \nu_{\mathrm{i}}^k, p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1} \rangle &\geq \langle \nabla F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}), p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1} \rangle \\
&= \langle -\nu_{\mathrm{i}}^{k+1}, p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1} \rangle + \langle \nu_{\mathrm{i}}^k, p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1} \rangle \\
&= -c \langle p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^k, p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1} \rangle.
\end{aligned}$$

Then by the identity (2.14):

$$c \langle p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^k, p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1} \rangle = \frac{c}{2} \left[\|p_{\mathrm{i}}^k - B_{\mathrm{i}} q^k\|^2 - \|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^k\|^2 - \|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2 \right].$$

Putting the above results together, we have

$$\begin{aligned}
&\mathcal{L}_c(\{p_{\mathrm{i}}^{k+1}\}_{<\mathrm{i}}, p_{\mathrm{i}}^k, \{p_{\mathrm{i}}^k\}_{>\mathrm{i}}, q^k, \nu_{\mathrm{i}}^k) - \mathcal{L}_c(\{p_{\mathrm{i}}^{k+1}\}_{<\mathrm{i}}, p_{\mathrm{i}}^{k+1}, \{p_{\mathrm{i}}^k\}_{>\mathrm{i}}, q^k, \nu_{\mathrm{i}}^k) \\
&\qquad = \frac{c}{2} \|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2 + \frac{c}{2} \|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^k\|^2 - \frac{c}{2} \|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^k\|^2. \quad (3.45)
\end{aligned}$$

Next, for the dual update:

$$\mathcal{L}_c(\{p_{\mathrm{i}}^{k+1}\}, q^k, \{\nu_{\mathrm{i}}^{k+1}\}_{<\mathrm{i}}, \nu_{\mathrm{i}}^k, \{\nu_{\mathrm{i}}^k\}) - \mathcal{L}_c(\{p_{\mathrm{i}}^{k+1}\}, q^k, \{\nu_{\mathrm{i}}^{k+1}\}_{<\mathrm{i}}, \nu_{\mathrm{i}}^k, \{\nu_{\mathrm{i}}^k\}) = -c\|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^k\|^2. \tag{3.46}$$

Lastly, for the q-update, we have:

$$\begin{aligned}\mathcal{L}_c(\{p_{\mathrm{i}}\}, q^k, \{\nu_{\mathrm{i}}^{k+1}\}) - \mathcal{L}_c(\{p_{\mathrm{i}}\}, q^{k+1}, \{\nu_{\mathrm{i}}^{k+1}\}) &= G(q^k) - G(q^{k+1}) \\ &+ \sum_{\mathrm{i}=1}^{S}\langle \nu_{\mathrm{i}}^{k+1}, B_{\mathrm{i}}q^{k+1} - B_{\mathrm{i}}q^k\rangle + \frac{c}{2}\sum_{\mathrm{i}=1}^{S}\left(\|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^k\|^2 - \|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1}\|^2\right).\end{aligned} \tag{3.47}$$

Then by the weak convexity of G:

$$G(q^k) - G(q^{k+1}) \geq \langle \nabla G(q^{k+1}), q^k - q^{k+1}\rangle - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2.$$

Observe that:

$$\begin{aligned}&\langle \nabla G(q^{k+1}), q^k - q^{k+1}\rangle + \sum_{\mathrm{i}=1}^{S}\langle \nu_{\mathrm{i}}^{k+1}, B_{\mathrm{i}}q^{k+1} - B_{\mathrm{i}}q^k\rangle \\ &= \sum_{\mathrm{i}=1}^{S}\langle B_{\mathrm{i}}\left[\nu_{\mathrm{i}}^{k+1} + c\left(p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1}\right)\right], q^k - q^{k+1}\rangle + \sum_{\mathrm{i}=1}^{S}\langle \nu_{\mathrm{i}}^{k+1}, B_{\mathrm{i}}q^{k+1} - B_{\mathrm{i}}q^k\rangle \\ &= c\sum_{\mathrm{i}=1}^{S}\langle p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1}, B_{\mathrm{i}}q^k - B_{\mathrm{i}}q^{k+1}\rangle \\ &= \frac{c}{2}\sum_{\mathrm{i}=1}^{S}\left[-\|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^k\|^2 + \|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1}\|^2 + \|B_{\mathrm{i}}q^k - B_{\mathrm{i}}q^{k+1}\|^2\right],\end{aligned} \tag{3.48}$$

where the last equality follows the identity (2.14). Substitute (3.48) into (3.47), along with the weak convexity of G, we get:

$$\begin{aligned}\mathcal{L}_c(\{p_{\mathrm{i}}\}, q^k, \{\nu_{\mathrm{i}}^{k+1}\}) - \mathcal{L}_c(\{p_{\mathrm{i}}\}, q^{k+1}, \{\nu_{\mathrm{i}}^{k+1}\}) &\geq -\frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 \\ &+ \frac{c}{2}\sum_{\mathrm{i}=1}^{S}\left[\|B_{\mathrm{i}}q^k - B_{\mathrm{i}}q^{k+1}\|^2 - \|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^k\|^2\right].\end{aligned} \tag{3.49}$$

Combining (3.45) (3.46) and (3.49), we arrive at:

$$
\begin{aligned}
&\mathcal{L}_c^k - \mathcal{L}_c^{k+1} \\
&\geq -\frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 + \frac{c}{2}\sum_{\mathrm{i}=1}^{S}\left[\|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2 + \|B_{\mathrm{i}}q^k - B_{\mathrm{i}}q^{k+1}\|^2\right] - \frac{1}{c}\sum_{\mathrm{i}=1}^{S}\|\nu_{\mathrm{i}}^k - \nu_{\mathrm{i}}^{k+1}\|^2 \\
&\geq \sum_{\mathrm{i}=1}^{S}\left(\frac{c}{2} - \frac{L_{\mathrm{i}}^2}{c}\right)\|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2 + \frac{c}{2}\|B_{\mathrm{i}}q^k - B_{\mathrm{i}}q^{k+1}\|^2 - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 \\
&\geq \sum_{\mathrm{i}=1}^{S}\left(\frac{c}{2} - \frac{L_{\mathrm{i}}^2}{c}\right)\|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2 + \sum_{\mathrm{i}=1}^{S}\left(\frac{\sigma_G}{2} - \frac{c}{2M_q^2}\right)\|q^k - q^{k+1}\|^2,
\end{aligned}
$$

where the second inequality is due to the L_{i}-smoothness of F_{i} and the minimizer condition (3.22a); whereas in the last inequality is due to the M_q-Lipschitz continuity of G and the sub-minimization path technique [45]. The make the coefficients pre-multiplied by the square norms be all non-negative and not all equal to zeros, we pick the penalty coefficient c^*:

$$
c^* > \max_{\mathrm{i}\in[S]}\{\sqrt{2}L_{\mathrm{i}}, \sigma_G M_q^2\},
$$

which is satisfied under Assumption E. □

Given the sufficient decrease lemma, the convergence of the MsPF algorithm follows. To further prove the Q-linear rate of convergence, we again need to show the augmented Lagrangian (3.20), solved with the algorithm (3.21) has the Łojasiewicz exponent $\theta = 1/2$. This can be achieved through the following two results. The first is an upper bound of the difference $\mathcal{L}_c(w^{k+1}) - \mathcal{L}_c(w^*)$.

Lemma 3.3.9. *Let $\mathcal{L}_c$ be defined as in (3.20). Suppose Assumption E is satisfied, then for the sequence $\{w^k\}_{k\in\mathbb{N}}$ obtained through the algorithm (3.21), with $w^k := (\{p_{\mathrm{i}}^k\}, q^k, \{\nu_{\mathrm{i}}^k\})_{\mathrm{i}=1}^S$ denotes the collective point at step k, we have:*

$$
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \sum_{\mathrm{i}=1}^{S}\left(\frac{\sigma_G M_q^2 - c}{2}\|B_{\mathrm{i}}q^{k+1} - B_{\mathrm{i}}q^*\|^2 + \frac{c}{2}\|p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}}^*\|^2\right),
$$

where $\mathcal{L}_c^k := \mathcal{L}_c(w^k)$ and w^* denotes a stationary point whose local neighborhood is such that for all $\|w - w^*\| < \varepsilon$, $\mathcal{L}_c^* < \mathcal{L}_c < \mathcal{L}_c^* + \delta$ for some constants $\varepsilon, \delta > 0$.

Proof. First, by the convexity of $F_{\mathrm{i}}, \forall \mathrm{i} \in [S]$:

$$F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}) - F_{\mathrm{i}}(p_{\mathrm{i}}) \leq \langle \nabla F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}), p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}} \rangle = -\langle \nu_{\mathrm{i}}^{k+1}, p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}} \rangle,$$

where the last equality is due to the minimizer condition (3.22a). Second, by the weak convexity of $G(q)$:

$$\begin{aligned} G(q^{k+1}) - G(q) &\leq \langle \nabla G(q^{k+1}), q^{k+1} - q \rangle + \frac{\sigma_G}{2} \|q^{k+1} - q\|^2 \\ &= \sum_{\mathrm{i}=1}^{S} \langle \nu_{\mathrm{i}}^{k+1} + c \left(p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^{k+1} \right) \rangle + \frac{\sigma_G}{2} \|q^{k+1} - q^k\|^2, \end{aligned}$$

where we substitute the gradient of G with the minimizer condition (3.22c) to obtain the last equality. Combining the above results, we get:

$$\begin{aligned} &\mathcal{L}_c(\{p_{\mathrm{i}}^{k+1}\}, q^{k+1}, \{\nu_{\mathrm{i}}^{k+1}\})_{\mathrm{i}=1}^{S} - \mathcal{L}_c(\{p_{\mathrm{i}}\}, q, \{\nu_{\mathrm{i}}\})_{\mathrm{i}=1}^{S} \\ \leq &\sum_{\mathrm{i}=1}^{S} \langle \nabla F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}), p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}} \rangle + \langle \nabla G(q^{k+1}), q^{k+1} - q \rangle + \frac{\sigma_G}{2} \|q^{k+1} - q\|^2 \\ &\quad + \sum_{\mathrm{i}=1}^{S} \left[\langle \nu_{\mathrm{i}}^{k+1}, p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^{k+1} \rangle - \langle \nu, p_{\mathrm{i}} - B_{\mathrm{i}} q \rangle \right] \\ &\quad + \frac{c}{2} \sum_{\mathrm{i}=1}^{S} \left(\|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^{k+1}\|^2 - \|p_{\mathrm{i}} - B_{\mathrm{i}} q\|^2 \right) \\ = &\sum_{\mathrm{i}=1}^{S} \langle \nu_{\mathrm{i}}^{k+1}, p_{\mathrm{i}} - B_{\mathrm{i}} q \rangle + c \sum_{\mathrm{i}=1}^{S} \langle p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^{k+1}, B_{\mathrm{i}} q^{k+1} - B_{\mathrm{i}} q \rangle + \frac{\sigma_G}{2} \|q^{k+1} - q\|^2 \\ &\quad + \frac{c}{2} \sum_{\mathrm{i}=1}^{S} \left(\|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^{k+1}\|^2 - \|p_{\mathrm{i}} - B_{\mathrm{i}} q\|^2 \right). \end{aligned}$$

Substitute a stationary point w^* into the above inequality. Note that at w^*, the relation $p_{\mathrm{i}}^* = B_{\mathrm{i}} q^*, \forall \mathrm{i} \in [S]$, we have the simplified expression:

$$\begin{aligned} &\mathcal{L}_c(\{p_{\mathrm{i}}^{k+1}\}, q^{k+1}, \{\nu_{\mathrm{i}}^{k+1}\})_{\mathrm{i}=1}^{S} - \mathcal{L}_c(\{p_{\mathrm{i}}^*\}, q^*, \{\nu_{\mathrm{i}}^*\})_{\mathrm{i}=1}^{S} \\ &\quad \leq c \sum_{\mathrm{i}=1}^{S} \langle p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^{k+1}, B_{\mathrm{i}} q^{k+1} - B_{\mathrm{i}} q^* \rangle + \frac{\sigma_G}{2} \|q^{k+1} - q^*\|^2 + \frac{c}{2} \sum_{\mathrm{i}=1}^{S} \|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^{k+1}\|^2. \end{aligned}$$

For the inner product in the above, by the identity (2.14), we have:

$$\begin{aligned}
&\mathcal{L}_c(\{p_{\mathrm{i}}^{k+1}\}, q^{k+1}, \{\nu_{\mathrm{i}}^{k+1}\})_{\mathrm{i}=1}^{S} - \mathcal{L}_c(\{p_{\mathrm{i}}^{*}\}, q^{*}, \{\nu_{\mathrm{i}}^{*}\})_{\mathrm{i}=1}^{S} \\
\leq &\frac{c}{2}\sum_{\mathrm{i}=1}^{S}\left[\|p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}}^{*}\|^2 - \|B_{\mathrm{i}}q^{k+1} - B_{\mathrm{i}}q^{*}\|^2\right] + \frac{\sigma_G}{2}\|q^{k+1} - q^{*}\|^2 \\
\leq &\sum_{\mathrm{i}=1}^{S}\left(\frac{\sigma_G M_q^2 - c}{2}\|B_{\mathrm{i}}q^{k+1} - B_{\mathrm{i}}q^{*}\|^2 + \frac{c}{2}\|p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}}^{*}\|^2\right),
\end{aligned}$$

where the last inequality is due to the M_q-smoothness of the function G along with the sub-minimization path technique [45]. □

The next lemma provides a relation of the square norm of the gradient $\nabla\mathcal{L}_c$, evaluated with a step $k+1$ solution.

Lemma 3.3.10. *Let $\mathcal{L}_c$ be defined as in (3.20). Suppose Assumption E is satisfied, then for the sequence $\{w^k\}_{k\in\mathbb{N}}$ where $w^k := (\{p_{\mathrm{i}}^k\}, q^k, \{\nu_{\mathrm{i}}^k\})_{\mathrm{i}=1}^{S}$ obtained through the algorithm (3.21), we have:*

$$\|\mathcal{L}_c(w^{k+1})\|^2 = (c^2+1)\sum_{\mathrm{i}=1}^{S}\|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1}\|^2.$$

Proof. By the definition of the augmented Lagrangian (3.20), we have for $\mathrm{i} \in [S]$:

$$\nabla\mathcal{L}_c^{k+1} = \begin{bmatrix} \nabla F_{\mathrm{i}}(p_{\mathrm{i}}^{k+1}) + \nu_{\mathrm{i}}^{k+1} + c\left(p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1}\right) \\ \nabla G(q^{k+1}) - \sum_{\mathrm{i}=1}^{S} B_{\mathrm{i}}^T\left[\nu_{\mathrm{i}}^{k+1} + c\left(p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1}\right)\right] \\ p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1} \end{bmatrix}.$$

Substitute the minimizer conditions (3.22) into the above, we have:

$$\nabla\mathcal{L}_c^{k+1} = \begin{bmatrix} c\left(p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1}\right) \\ 0 \\ p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1} \end{bmatrix}.$$

Hence, $\|\nabla\mathcal{L}_c^{k+1}\|^2 = (c^2+1)\|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}}q^{k+1}\|^2$. □

Then, by combining Lemma (3.3.9) and Lemma 3.3.10, we can prove that the Łojasiewicz exponent $\theta = 1/2$ for the MsPF Lagrangian.

Lemma 3.3.11. *Let $\mathcal{L}_c$ be defined as in (3.20). Suppose Assumption E is satisfied and the sequence $\{w^k\}_{k\in\mathcal{N}}$ is obtained from the algorithm (3.21), where $w^k := (\{p_{\mathrm{i}}^k\}, q^k, \{\nu_{\mathrm{i}}^k\})_{\mathrm{i}=1}^S$ denotes the collective point at step k, then the Łojasiewicz exponent of the augmented Lagrangian $\mathcal{L}_c$ is $\theta = 1/2$ around a local neighborhood of a stationary point w^* such that $\|w - w^*\| < \varepsilon$ and $\mathcal{L}_c^* < \mathcal{L}_c < \mathcal{L}_c^* + \delta$ for some constants $\delta, \varepsilon > 0$.*

Proof. By Lemma 3.3.9, we have:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* &\leq \sum_{\mathrm{i}=1}^S \left(\frac{\sigma_G M_q^2 - c}{2} \|B_{\mathrm{i}} q^{k+1} - B_{\mathrm{i}} q^*\|^2 + \frac{c}{2} \|p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}}^*\|^2 \right) \\
&\leq \frac{c}{2} \sum_{\mathrm{i}=1}^S \|p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}}^*\|^2 \\
&\leq \frac{c}{2} \sum_{\mathrm{i}=1}^S \left(\|p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}}^*\|^2 + \|p_{\mathrm{i}}^{k+1} - B_{\mathrm{i}} q^{k+1}\|^2 \right),
\end{aligned}$$

where the second inequality is due to the Assumption E, where $c > \max_{\mathrm{i}\in[S]}\{\sqrt{2}L_{\mathrm{i}}, \sigma_G M_q^2\}$ and hence the first term is negative. Then by Lemma 3.3.10, we have an upper bound of the above inequality:

$$\begin{aligned}
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* &\leq \frac{c}{2} \sum_{\mathrm{i}=1}^S \left(\|p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}}^*\|^2 + \frac{1}{1+c^2} \|\nabla \mathcal{L}_c^{k+1}\|^2 \right) \\
&= \frac{c}{2} \|\nabla \mathcal{L}_c^{k+1}\|^2 \sum_{\mathrm{i}=1}^S \left(\frac{1}{1+c^2} + \frac{\|p_{\mathrm{i}}^{k+1} - p_{\mathrm{i}}^*\|^2}{\|\nabla \mathcal{L}_c^{k+1}\|^2} \right) \\
&\leq \frac{c}{2} \|\nabla \mathcal{L}_c^{k+1}\|^2 \left(\frac{S}{1+c^2} + \frac{\|w^{k+1} - w^*\|^2}{\|\nabla \mathcal{L}_c^{k+1}\|^2} \right) \\
&\leq \frac{c}{2} \|\nabla \mathcal{L}_c^{k+1}\|^2 \left(\frac{S}{1+c^2} + \frac{\varepsilon^2}{\xi^2} \right),
\end{aligned}$$

where the last inequality is due to assumption that there the existence of a local neighborhood around w^*, which implies that $\eta > \|\nabla \mathcal{L}_c^{k+1}\| > \xi$ for some small constants $\eta, \xi > 0$, by Lemma 2.4.11. Then by taking square root on the both sides of the last inequality, we complete the proof. □

Finally, since the Łojasiewicz exponent $\theta = 1/2$, we can apply Lemma 2.4.10 which implies the corresponding rate of convergence is Q-linear, the desired result. We formulate this result into the following theorem.

Theorem 3.3.12. *Let the augmented Lagrangian $\mathcal{L}_c$ defined as in (3.20). Suppose Assumption E is satisfied, and the sequence $\{w^k\}_{k\in\mathbb{N}}$ is obtained through the algorithm (3.21) where $w^k := (\{p_{\mathrm{i}}^k\}, q^k, \{\nu_{\mathrm{i}}^k\})_{\mathrm{i}=1}^S$ denotes the collective point at step k, then $\{w^k\}$ converges to a stationary point w^* at Q-linear rate around the local neighborhood such that $\|w - w^*\| < \varepsilon$ and $\mathcal{L}_c^* < \mathcal{L}_c < \mathcal{L}_c + \delta$ for some constants $\varepsilon, \delta > 0$.*

Proof. The proof consists of two parts, first we show that the sequence $\{w^k\}_{k\in\mathbb{N}}$ is convergent toward a stationary point w^*. Then we adopt the KŁ inequality to prove the rate of convergence is Q-linear locally around w^*. For the convergence, under Assumption E, and the sufficient decrease lemma (Lemma 3.3.8), we have:

$$\mathcal{L}_c^0 - \mathcal{L}_c^N \geq \sum_{k=0}^{N-1} \mathcal{L}_c^k - \mathcal{L}_c^{k+1} \geq K^* \sum_{k=1}^{N-1} \left(\sum_{\mathrm{i}=1}^{S} \|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\|^2 + \|q^k - q^{k+1}\|^2 \right),$$

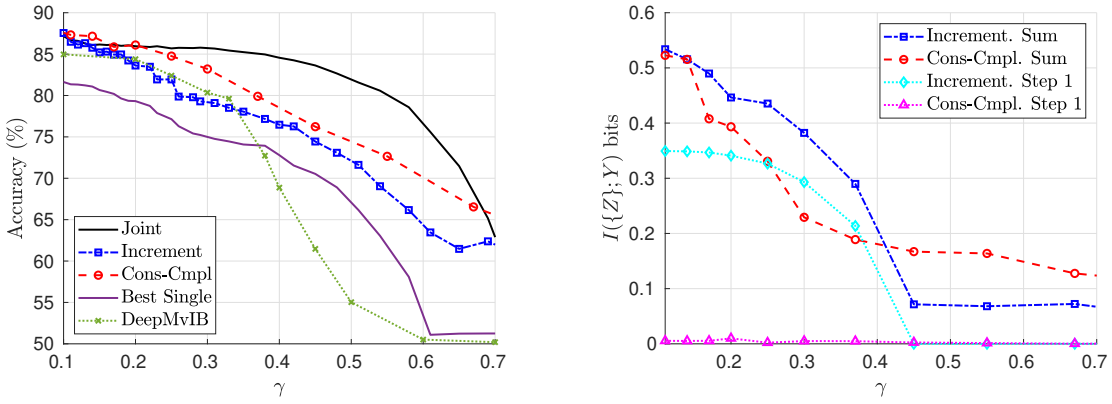
for some constant $K^* > 0$ such that the penalty coefficient c satisfies the Assumption E. Then since $\mathcal{L}_c$ is lower semi-continuous due to the discrete setting, the left-hand-side of the above inequality is finite $\forall N \in \mathbb{N}$. Then, observe that the right-hand-side of the inequality above is a Cauchy sequence, and hence is bounded. In turns, we have $\forall \mathrm{i} \in [S], \|p_{\mathrm{i}}^k - p_{\mathrm{i}}^{k+1}\| \to 0, \|q^k - q^{k+1}\| \to 0$ as $k \to \infty$. Moreover, due to L_{i}-smoothness and the minimizer condition (3.22b), $\|\nu_{\mathrm{i}}^k - \nu_{\mathrm{i}}^{k+1}\| \to 0, \forall \mathrm{i} \in [S]$. This implies the convergence toward a stationary point that belongs to a critical set $\Omega^* := \{w | \nabla \mathcal{L}_c(w) = 0, \forall w \in \Omega\}$ where Ω is the set of all the feasible solutions. In other words, we have $w^k \to w^* \in \Omega^*$ as $k \to \infty$. This proves the convergence of the sequence $\{w^k\}$ for $k > N_0 \in \mathbb{N}$ sufficiently large. Then to apply the KŁ

inequality, we need to establish the relation: $\|\nabla \mathcal{L}_c^k\| \leq R^* \|w^k - w^{k-1}\|$ for some constant $R^* > 0$. By Lemma 3.3.10, we have:

$$\begin{aligned}
\|\nabla \mathcal{L}_c^k\|^2 &= (c^2+1) \sum_{\mathrm{i}=1}^{S} \|p_{\mathrm{i}}^k - B_{\mathrm{i}} q^k\|^2 \\
&\leq \left(c^2+1\right) \sum_{\mathrm{i}=1}^{S} \|p_{\mathrm{i}}^k - B_{\mathrm{i}} q^{k-1} + B_{\mathrm{i}} q^{k-1} - B_{\mathrm{i}} q^k\|^2 \\
&\leq 2\left(c^2+1\right) \sum_{\mathrm{i}=1}^{S} \left(\|p_{\mathrm{i}}^k - B_{\mathrm{i}} q^{k-1}\|^2 + \|B_{\mathrm{i}} q^{k-1} - B_{\mathrm{i}} q^k\|^2\right) \\
&= 2\left(c^2+1\right) \sum_{\mathrm{i}=1}^{S} \left(\frac{1}{c^2}\|\nu_{\mathrm{i}}^k - \nu_{\mathrm{i}}^{k-1}\|^2 + \lambda_{B_{\mathrm{i}}^T B_{\mathrm{i}}} \|q^k - q^{k-1}\|^2\right) \\
&\leq 2(c^2+1) \lambda_{B^T B}^* \|w^k - w^{k-1}\|^2,
\end{aligned}$$

where we apply Cauchy-Schwarz inequality to obtain the second inequality, followed by the dual update $\nu_{\mathrm{i}}^k = \nu_{\mathrm{i}}^{k-1} + c(p_{\mathrm{i}}^k - B_{\mathrm{i}} q^k)$; in the last inequality we define $\lambda_{B^T B}^* := \max_{\mathrm{i} \in [S]} \{1/c^2, \lambda_{B_{\mathrm{i}}^T B_{\mathrm{i}}}\}$ with λ_M denotes the largest eigenvalue of a matrix M. Then taking square root of both sides of the above inequality, we obtain the desired relation. Then along with Lemma 3.3.11, which gives $\theta = 1/2$, we can apply Lemma 2.4.10 and prove the corresponding rate of convergence is Q-linear. □

As a remark, Theorem 3.3.7 and Theorem 3.3.12 demonstrate that the rates of convergence of the MvIB (3.11) and MsPF (3.21) algorithms are both Q-linear. On the other hand, while Theorem 3.3.3 only implies R-linear rate of convergence, we hypothesized that the R-linear result is needed if we further generalize the algorithm to DRS methods, where a relaxation parameter $\alpha > 0$ is included, and a similar division of the region of linear rate of convergence can be shown just as in the single modal cases. We leave this direction for future exploration.



(a) Classification Accuracy (b) Mutual information

Figure 3.1. Simulation results of the proposed and compared MvIB methods on synthetic datasets

3.4 Evaluation

3.4.1 Synthetic Data: Classification Task

We evaluate the proposed two approaches with a synthetic dataset consisting of two views of distributions. For convenience of expression, we denote the consensus-complement approach as *Cons-Cmpl* while the incremental update approach as *Increment*.

We implement the the Bayes' decoder for *Cons-Cmpl* as:

$$p_{cc}(y|x^{(1)},x^{(2)}) = \sum_{\{z_e^{(1)},z_e^{(2)},z_c^{(1)},z_c^{(2)}\}} p(y|z_c^{(1)},z_c^{(2)},z_e^{(1)},z_e^{(2)})p(z_e^{(1)},z_c^{(1)}|x^{(1)})p(z_e^{(2)},z_c^{(2)}|x^{(2)}),$$

On the other hand, the Bayes' decoder for *Increment* is implemented as:

$$p_{inc}(y|x^{(1)},x^{(2)}) = \sum_{\{z^{(1)},z^{(2)}\}} p(y|z^{(1)},z^{(2)})p(z^{(1)}|x^{(1)})p(z^{(2)}|z^{(1)},x^{(2)}).$$

For the parameters, we set $c = 64, \min\{\varepsilon\} = 10^{-11}$ and run the algorithms with random initialization. For simplicity, we let $\gamma_1 = \gamma_2 = \gamma$. The termination criterion for both the proposed algorithms is either when the total variation, i.e., the linear constraints, between the primal and augmented variables $D_{TV}(A_i p_i||q) < 10^{-6}, \forall i \in [V]$ (convergent case), or the maximum number of iteration is reached (divergent case). Figure 3.1a follows the distribution

given in (3.50) with the label *Joint* denoting the joint-view IB approach as a compared method; The distribution in Figure 3.1b is given in (3.51).

We simulate a classification task and compare the performance of the two proposed approaches to joint-view/single-view IB solvers [22]. The compared methods served as references for the best and the worst case performance. We also compare the state-of-the-art deep neural network-based method (DeepMvIB [9]), where we implemented with two layers of 4-neuron, fully connected weights plus ReLU activation for each view. In generating the data samples, given the joint distribution (3.50), we randomly sample 10000 pairs of outcomes $(y, x^{(1)}, x^{(2)})$ as testing data. Then we run the algorithms, sweeping through a range of $\gamma \in [0.1, 0.7]$ and record the best accuracy for each approach from 50 trials per γ. We use Bayes' decoder to predict the testing data, where we perform inverse transform sampling for the cumulative distribution of the decoders to obtain $\hat{y}$ for each pair of $(x^{(1)}, x^{(2)})$. The data-generating distribution is given as:

$$P(X^{(1)}|Y) = \begin{bmatrix} 0.75 & 0.05 \\ 0.20 & 0.20 \\ 0.05 & 0.75 \end{bmatrix}, P(X^{(2)}|Y) = \begin{bmatrix} 0.85 & 0.15 \\ 0.15 & 0.85 \end{bmatrix}, \tag{3.50}$$

where $P(Y) = \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}^T$. The result is shown in Figure 3.1a. The dimension of each of $Z_c, Z_e^{(2)}, Z^{(2)}$ is 2, and 3 for each of $Z_e^{(1)}, Z^{(1)}$. The figure clearly shows that the two proposed approaches can achieve comparable performance to that of the joint-view IB solver, which served as the best performance reference. Moreover, both our solvers outperform the deepMvIB over the range of the trade-off parameter γ we simulated. Interestingly, in comparing the two proposed solvers, *Cons-Cmpl* outperforms *Increment* in the best accuracy achieved. Intuitively, we hypothesize this might be due to the abundance of representation overlap within the two joint distributions (3.50). To better investigate this hypothesis, we

further consider a different set of distributions with dimensions of all representations given as $|\mathcal{Z}_c| = |\mathcal{Z}_{\mathrm{e}}^{(\mathrm{i})}| = |\mathcal{Z}^{(\mathrm{i})}| = 3, \forall \mathrm{i} \in \{1, 2\}$:

$$p(Y) = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}^T, P(X^{(1)}|Y) := \begin{bmatrix} 0.90 & 0.20 & 0.20 \\ 0.05 & 0.45 & 0.35 \\ 0.05 & 0.35 & 0.45 \end{bmatrix},$$
$$P(X^{(2)}|Y) := \begin{bmatrix} 0.25 & 0.10 & 0.55 \\ 0.20 & 0.80 & 0.25 \\ 0.55 & 0.10 & 0.20 \end{bmatrix}. \tag{3.51}$$

Observe that for each view in (3.51), there is one class (y_1 in view 1 and y_2 in view 2), that is easy to infer through $X^{(\mathrm{i})}, \mathrm{i} \in \{1, 2\}$ while the remaining two are ambiguous. This results in low representation overlap and hence, consensus is difficult to form. In Figure 3.1b we examine the components of the relevance rate $I(\{Z\}; Y)$ where the *Sum* is: $I(Z_c; Y) + \sum_{\mathrm{i}=1}^{2} I(Z_{\mathrm{e}}^{(\mathrm{i})}; Y|Z_c))$ for *Cons-Cmpl* and $I(Z^{(1)}; Y) + I(Z^{(2)}; Y|Z^{(1)})$ for *Increment*. *Step 1* indicates $I(Z_c; Y)$ in *Cons-Cmpl*, and $I(Z^{(1)}; Y)$ in *Increment*. Observe that there is almost no increase in $I(Z_c; Y)$ over varying γ, and that *Increment* has a greater relevance rate than *Cons-Cmpl* when $\gamma < 0.4$. Since it is known that the high relevance rate is related to high prediction accuracy [15], this example favors the *Increment*. In interpreting this result, since *Increment*, by design, increases the overall relevance rate view-by-view so it does not form a consensus from all views in one instance, it works well in the case where there is limited representation overlap. However, in the opposite case, as we demonstrated in the first simulation, *Cons-Cmpl* can be a more advantageous algorithm.

3.4.2 Real-World Data: Feasibility for Large-Scale Problems

To show that the proposed framework can apply to real-world data and demonstrate its feasibility. We use the celebrated MNIST dataset [81], consisting of hand-written digits. In MNIST, each data sample is a 28×28 pixels grayscale digit with a ground-truth label indicating what the digit represents. We pre-process the dataset into two views by clipping the upper-half of an image (14×28 pixels) as the first view and the lower-half of the image

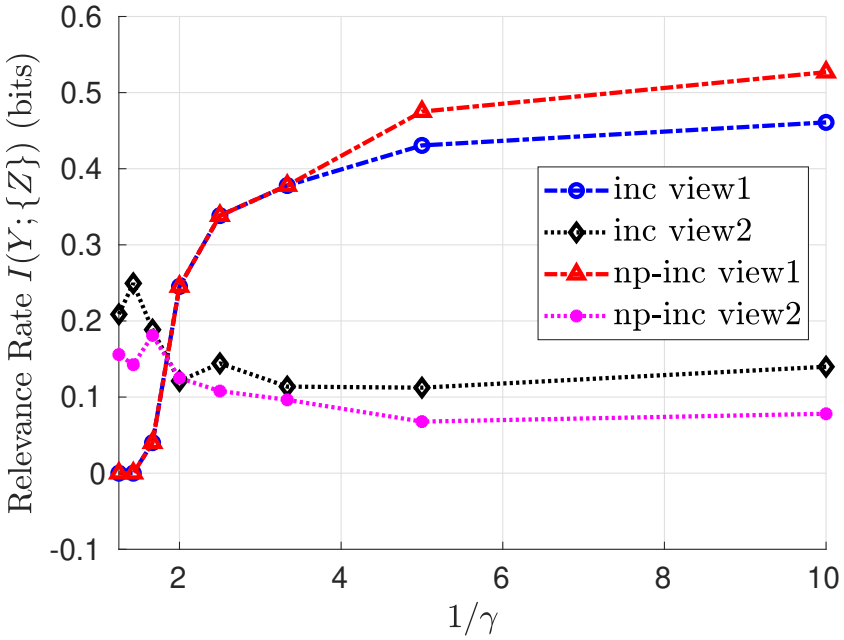


Figure 3.2. Comparing NumPy and Tensorflow Implementations

(14×28 pixels) as the second view. Instead of using all 10 digits, we select three of them: $[1, 6, 8]$ of the testing dataset of MNIST which corresponds to 3067 instances. For the joint probability, we fit two deep variational IB (VIB [2]) for each of the two views of partial digits respectively. Both VIB models attain above 96% testing accuracy. Then for the partial digit in each view, we feed the images to the corresponding VIB model which give two sets of soft prediction $\{\hat{p}_\theta(y|x_{\mathrm{i}}^{(1)}), \hat{p}_\theta(y|x_{\mathrm{i}}^{(2)})\}$. Then by collecting all $\mathrm{i} \in [N]$ prediction vectors into a matrix $P_{Y|X^{(1)}}$ whose columns are the prediction vectors, and assuming uniform occurrence of each view observation $p(x^{(1)}) = 1/|X^{(1)}|, p(x^{(2)}) = 1/|X^{(2)}|$ we obtain the joint probability. Before presenting the result, we provide some implementation details to support the feasibility of the proposed methods. To deal with the significant increase of observation dimensions and maintain correctness on small-scale toy problem as in Section 3.4.1, we implement both the consensus-complement and incremental update models on both Tensorflow 2 [82] and NumPy [83] exploiting the large-scale capability of the former and the easy-to-implement feature of the latter. Figure 3.2 compares the two models on the toy problem in Section 3.4.1 for the incremental-update method, similar results hold

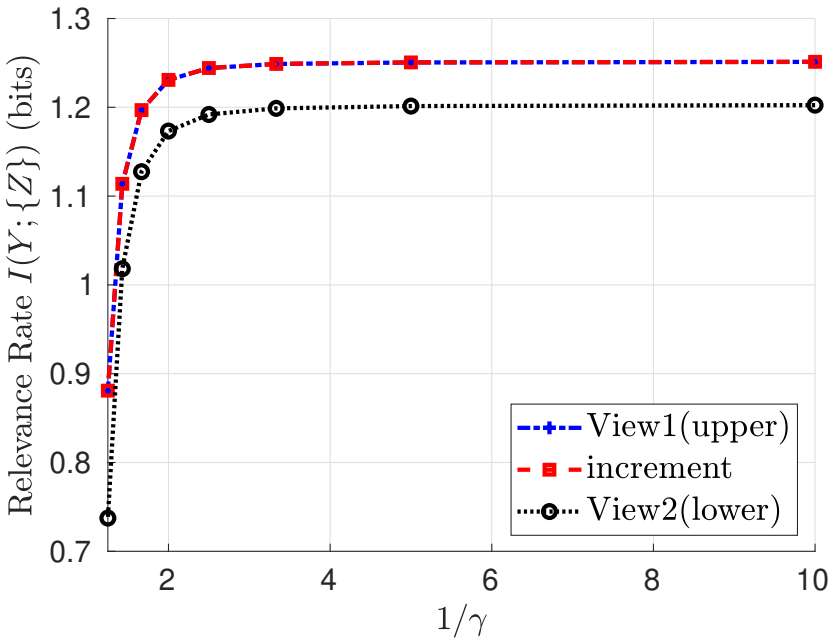


Figure 3.3. Relevance rate of the two-view MNIST predictions

for the consensus-complement model. We show the step-wise increase of the relevance rates for the two implementations and the results verify that the conversion of the program to different environments is successful. Then in Figure 3.3, we present the modified multi-view MNIST evaluation. For comparison purposes, two single-view results are shown along with the one for the increment-update. Interestingly, the relevance rate cannot be improved with multi-view observations, this is because even with partial digits, the accuracy for each view is reasonably high ($> \%96$), so there is no more representation overlap to exploit. In other words, as there is no further improvement that can be made, the highest possible relevance rate is dominated by the one that is most informative (view 1), and the incremental-update method can achieve this value over a wide range of configurations.

3.4.3 Asymptotic Complexity

Finally, to demonstrate the complexity reduction from the proposed approaches, we compare the complexity of the two approaches in terms of the number of dimensions for the primal variables. For simplicity, let $|X| = |X^{(\mathrm{i})}|$,$|Z| = |Z_{\mathrm{c}}| = |Z_{\mathrm{e}}| = |Z^{(\mathrm{i})}|$. For *Cons-*

Cmpl, the number of dimensions for the variables scales as $\mathcal{O}(V|X||Z|^2)$ while for *Increment*, it grows as $\mathcal{O}(|X||Z|^V)$. In MvIB literature, it is more desirable to find a low-dimensional representation, $|X| \gg |Z|$ in practice. Based on this common practice, the two methods both improve over the joint view as their complexity values scale as $\mathcal{O}(|X|^V|Z|)$. Remarkably, the complexity for *Cons-Cmpl* scales linearly in the number of views V while we get an exponential growth with factor $|Z|$ for *Increment*. This complexity gain extends to the case for MsPF and demonstrates the benefits of both the proposed formulation and the developed algorithms.

4. GENERALIZATION ERROR ANALYSIS UNDER DISTRIBUTION MISMATCH

In this chapter, we present a new information-theoretic perspective of the generalization error analysis, by viewing it as distribution mismatch. Different from previous works in this topic that derive the error bound from bounding the generalization error gap between the empirical risk and population risk, and then aiming at finding the theoretic performance guarantee in terms of the number of data, or the cardinality of the space of learning models within certain confidence level, or both, we formulate the generalization error problem into a minimax framework and derive tighter bounds than existing ones [14], [17], [18], [56].

When viewing the problem as distribution mismatch, we find that our framework can be applied to both the standard learning tasks, where the source of distribution mismatch simply comes from the finite sampling process, and the recent adversarial learning scenario, where an adversary have access to either the training data or learning model to create another source of distribution mismatch in addition to sampling mismatch [58], [84].

In solving the new framework, we derive surrogate error upper bounds based on the Pythagorean theorem which turns out to connect the recent input-output mutual information bounds and tighten the results therein. In addition to this finding, we further show that the proposed minimax framework connects to the IB methods for learning problems, which recently attracts significant attention from machine learning and data science research. We show through the strong data-processing inequality (SDPI [62], [85], [86]) that in minimizing the surrogate loss upper bound, there is a close relationship between finding the minimum of the generalization error in the proposed framework and the predictive IB Lagrangian known in literature [87].

4.1 Preliminaries

In this part, some definitions that will facilitate the discussion in the rest of this chapter will be stated. In literature, the mutual information-based bounds are based on the Donsker-

Varadhan's representation of the KL divergence, which can be shown through a broader class of divergence known as the f-divergence [62].

Definition 4.1.1. *Let $(\mathcal{X}, \mathcal{F})$ be a measure space where the two probability distributions P, Q belongs to and $f : \mathcal{X} \mapsto \mathbb{R}$ is a measurable function. Then the Donsker-Varadhan's representation of the KL divergence is:*

$$\sup_{f:\mathcal{X}\mapsto\mathbb{R}} E_P[f] - \log E_Q[\mathrm{e}^f] = D_{KL}[P||Q].$$

On the other hand, the sub-Gaussian assumption holds in a wide range of loss functions used in practice. For example, in supervised learning context, a common choice of the loss function is the zero-one loss, as it accounts for performance metrics such as average accuracy rate in discriminative and classification tasks. And since its range is bounded over $[0, 1]$ its variance is bounded by $1/4$ hence $1/2$-sub-Gaussian.

Definition 4.1.2. *A random variable X is said to be σ-sub-Gaussian if:*

$$\log E[\exp\{\lambda X\}] \leq \frac{\lambda\sigma^2}{2}, \quad \forall\lambda \in \mathbb{R}.$$

The sub-Gaussian property is commonly assumed in studying concentration inequalities. Some of the elementary inequalities in learning theory will be listed next.

Definition 4.1.3. *For a non-negative random variable X such that $X \leq 1$, the Chernoff inequality holds with a probability of at least $1 - \delta$:*

$$P\left(X > \log\frac{1}{\delta}\right) \leq E[\mathrm{e}^x]\delta.$$

The above definition is a variant of its typical form in probability theory, which can be easily shown by the relation $E[\mathrm{e}^{sY}]/\mathrm{e}^t \leq \mathrm{e}^{-t} = \delta$ for a negative-valued random variable Y and some $s > 0$ such that $X = \mathrm{e}^{sY}$. The Chernoff inequality introduces an additional parameter s, which serves as an exponent to capture the rate of concentration. By maximizing it, one could obtain a tighter bound.

In addition to this result, if a random variable is also sub-Gaussian and has a bounded range, then a variety of concentration inequalities applies and hence allowing one to establish generalization error bounds.

Another line of research in studying tight generalization error bounds that also starts with sub-Gaussian assumption for the loss functions and further shows that if the cumulant generating function (CGF) of a random variable X is bounded above by a class of finite-range functions, then it is possible to obtain tight upper bound to the probability when a deviation event occurs.

The above-mentioned approaches, when incorporated with the input-output mutual information, are recently found to be special cases of a unified framework, by introducing the α-information.

Definition 4.1.4. *For $\alpha \in (0,1) \bigcup (1,\infty)$, the α information of two Borel-measurable p, q with a compact support, the α-information of p and q is denotes as $D_\alpha(p||q)$ and is defined as:*

$$D_\alpha(p||q) := \frac{1}{\alpha - 1} \log \int p^\alpha q^{1-\alpha} d\mu.$$

4.2 Problem Formulation

In this part, we consider the following generalization error problem: Let $p_\varepsilon(x, y)$ denote an unknown joint distribution at testing phase with the random variables for the observations are denoted as X while the target variable is denoted as Y. On the other hand, we denote $p_\phi(x, y)$ the joint distribution for training that is available at training phase. A learner is given access to $p_\phi(x, y)$ (training distribution) and knows that the testing distribution $p_\varepsilon(x, y)$ is within a bounded divergence to $p_\phi(x, y)$, say $D_{KL}[p_\varepsilon(x, y)||p_\phi(x, y)] \leq \alpha$. The learner then develops an algorithm A which fits the training distribution to a desired level with a class of learning model, which is parameterized with θ, equivalently, the learned model can be expressed as a joint distribution $p_\theta(x, y)$. The goal of the learner is to pick an optimal fitting level, and hence θ^*, through a surrogate loss function that capture the partial knowledge at testing phase, hoping that the learned model will achieve minimal the distribution mismatch with respect to $p_\varepsilon(x, y)$.

The formulation can be expressed as the following minimax problem:

$$\theta^* = \arg\min_{\theta \in \Theta(\phi, A)} \max_{\varepsilon \in \mathcal{E}(\alpha, \phi)} D_{KL}[p_\varepsilon(x, y) || p_\theta(x, y)], \tag{4.1}$$

where the testing phase distribution mismatch $\mathcal{E}(\alpha, \phi)$ is defined as:

$$\mathcal{E}(\alpha, \phi) := \{\varepsilon | D_{KL}[p_\varepsilon(x, y) || p_\phi(x, y)] \leq \alpha, \varepsilon \in (\mathcal{X} \times \mathcal{Y}, \mathcal{F})\}, \tag{4.2}$$

where $(\mathcal{X} \times \mathcal{Y}, \mathcal{F})$ denote a measure space of ε. Whereas the set of learning model is defined as:

$$\Theta(\phi, \mathcal{A}) := \{\theta | \theta \in \mathcal{A}(\phi)\}, \tag{4.3}$$

where the algorithm $\mathcal{A}$ depends on a training distribution ϕ. Note that (4.1) includes both the standard learning theoretic settings, where the distribution mismatch comes from sampling solely, and the adversarial machine learning where an adversary crafted a joint distribution intended to worsen the performance of a learned model. To see this, if the set $\mathcal{E}(\alpha, \phi)$ contains a single element, corresponds to the unknown data-generating distribution then (4.1) recovers the standard learning theoretic setting; To see that (4.1) can include adversarial learning settings, consider from a learner's two-step decision process in solving the proposed minimax problem as follows. At the first step, the learner assumes the worst-case scenario with the indirect partial knowledge to the training data (i.e., $D_{KL}[p_\varepsilon(x, y) || p_\phi(x, y)] \leq \alpha$). Then at the second step, the learner picks the optimal learning model θ^* with desired fitting level to the training distribution $p_\phi(x, y)$, which equivalently, gives the joint distribution $p^*_\theta(x, y)$, believing that it can withstand the worst attack from a potential adversary whose power is limited to $\mathcal{E}(\alpha, \phi)$. Then clearly, the inner maximization problem and the outer minimization problem together fall within the context of (4.1).

Remarkably, the formulation connects to the formulation of the so-called input-output mutual information bounds [17], [18] where the input corresponds to ϕ and the output W, representing a hypothesis is generated from a stochastic kernel $p(W|\phi)$. Compared to the proposed framework, the stochastic kernel corresponds to (4.3) but can be thought of as a "maxmini" problem instead. Comparison to the approach will be deferred to Chapter 4.4.3.

4.3 Main Results

Our main results are the incorporation of side information about distribution mismatch to tighten the existing mutual information bounds. Moreover, we generalize the developed techniques to the α-information-based approaches which includes a broader class of information divergence.

Theorem 4.3.1. *Let $p_\varepsilon(x,y), p_\phi(x,y)$ be two Borel-measurable density functions defined in a measure space $S := (\mathcal{X}\times\mathcal{Y}, \mathcal{F})$. Given $p_\phi(x,y)$, define the following convex set, for $\xi \in S$:*

$$E(\alpha,\xi) := \{\varepsilon : S \mapsto \mathbb{R}, D_{KL}[p_\varepsilon(x,y)||p_\xi(x,y)] \leq \alpha\},$$

then we have:

$$D_{KL}[p_\varepsilon(x,y)||p_\phi(x,y)] \leq M(\varepsilon||\phi).$$

where:

$$M(\varepsilon||\phi) := \max\{m(\varepsilon,\phi), m(\phi,\varepsilon)\},$$

and $m(\varepsilon,\phi) := I_\varepsilon(X;Y) - I_\phi(X;Y) + D_{KL}[p_\varepsilon(x)||p_\phi(x)] + D_{KL}[p_\varepsilon(y)||p_\phi(y)]$.

Proof. First, due to the asymmetry of KL divergence, we have:

$$D_{KL}[p_\varepsilon(x,y)||p_\phi(x,y)] \leq \max\left\{D_{KL}[p_\varepsilon(x,y)||p_\phi(x,y)], D_{KL}[p_\phi(x,y)||p_\varepsilon(x,y)]\right\}.$$

Then without loss of generality, assume that $D_{KL}[p_\varepsilon(x,y)||p_\phi(x,y)] > D_{KL}[p_\phi(x,y)||p_\varepsilon(x,y)]$, then consider the following:

$$\begin{aligned}
D_{KL}[p_\varepsilon(x,y)||p_\phi(x,y)] &= \sum_{x,y} p_\varepsilon(x,y)\log\frac{p_\varepsilon(x,y)}{p_\phi(x,y)} \\
&= \sum_{x,y} p_\varepsilon(x,y)\log\frac{p_\varepsilon(x,y)}{p_\varepsilon(x)p_\varepsilon(y)}\frac{p_\varepsilon(x)p_\varepsilon(y)}{p_\phi(x)p_\phi(y)}\frac{p_\phi(x)p_\phi(y)}{p_\phi(x,y)} \\
&= I_\varepsilon(X;Y) + D_{KL}[p_\varepsilon(x)||p_\phi(x)] + D_{KL}[p_\varepsilon(y)||p_\phi(y)] \\
&\qquad + \sum_{x,y} p_\varepsilon(x,y)\log\frac{p_\phi(x)p_\phi(y)}{p_\phi(x,y)} \\
&\leq I_\varepsilon(X;Y) - I_\phi(X;Y) + D_{KL}[p_\varepsilon(x)||p_\phi(x)] + D_{KL}[p_\varepsilon(y)||p_\phi(y)]
\end{aligned}$$

where in the last inequality we apply the Pythagorean theorem [65]:

$$D_{KL}[P||Q] \geq D_{KL}[P||P^*] + D_{KL}[P^*||Q],$$

where $P^* := \inf_{P \in E} D_{KL}[P||Q]$ and we substitute $P = p_\varepsilon(x, y), Q = p_\phi(x)p_\theta(y)$ and $P^* = p_\theta(x, y)$. The proof is complete by defining a convex set E. It is straightforward to show that $E(\alpha, \xi)$ in the statement is convex, two joint probability $\varepsilon_1, \varepsilon_2 \in E(\alpha, \xi)$ and $\lambda \in [0, 1]$:

$$\alpha \geq \lambda D_{KL}[\varepsilon_1||\xi] + (1 - \lambda) D_{KL}[\varepsilon_2||\phi] \geq D_{KL}[\varepsilon_\lambda||\xi] \in E(\alpha, \xi),$$

where $\varepsilon_\lambda := \lambda\varepsilon_1 + (1-\lambda)\varepsilon_2$ and the second inequality follows the convexity of KL divergence. Note that while in deriving the result, we limit the focus on discrete random variables for simplicity, but the result can be easily generalized to continuous settings by substituting the summation with integration. □

Remarkably, by setting the joint probability $p_\phi(x, y) = p_\varepsilon(x)p_\varepsilon(y)$, that is, making it the product measure, then Theorem 4.3.1 recovers the recent input-output mutual information generalization error upper bounds [17], [18].

The key step in deriving Theorem 4.3.1 is by the Pythagorean theorem. Our goal is to demonstrate that the technique we proposed can be easily applied to various existing bounds or tightening techniques and therefore achieve improvements in the application of each method. To provide some examples, in Chapter 4.4, we apply Theorem 4.3.1 to a variety of existing mutual information-based bounds. Moreover, we find that the proposed technique can provide insights to the recent success of representation learning though the IB methods. Furthermore, beyond combining with existing approaches, our results can be applied to adversarial learning scenarios that recently gain significant attention as the discovery of adversarial samples in deep neural networks and the increasing concern over privacy leakage in data collection processes. In addition to theoretic results, in Chapter 4.5, we provide empirical evaluation of the proposed approaches to support the derived theoretical generalization error upper bounds. Interestingly, we find that our bounds are tighter than the existing ones due to the better exploitation of the training distribution.

4.4 Applications

4.4.1 Tight Upper Bound for Standard Learning Problem

In this part, we apply Theorem 4.3.1 to standard learning theoretic scenario, where the source of the distribution mismatch is sampling process solely. As mentioned in Chapter 4.2, the minimax problem (4.1) includes the standard learning theoretic scenario as a special case, here we explicitly describe the setup.

Consider the scenario where there is an unknown true data-generating joint distribution $p_\varepsilon(x, y)$ and a sampled version of the joint distribution $p_\phi(x, y)$, estimated from a finite training dataset of m samples, and is denoted as $S_m := \{z_\mathrm{i}\}_{\mathrm{i}=1}^m, z_\mathrm{i} = (x_\mathrm{i}, y_\mathrm{i}) \sim p_\varepsilon(x, y), \forall \mathrm{i} \in [m]$. A learner develops an algorithm A to optimize a learning model, parameterized by $\theta \in \Theta(A)$ through a surrogate loss function $\mathcal{L}(S_m, \theta) = D_{KL}[p_S(x, y)||p_\theta(x, y)]$, equivalently, the learned model can be expressed as another joint distribution $p_\theta(x, y)$. Given these resources, the goal of the learner is hoping to find the optimal $p^*_\theta(x, y)$ with minimum distribution shift with respect to the true joint distribution $p_\varepsilon(x, y)$, expressed as:

$$\theta^* := \arg\min_{\theta \in \Theta(A)} D_{KL}[p_\varepsilon(x, y)||p_\theta(x, y)]. \tag{4.4}$$

Note that in the above setup, since $E(\alpha)$ only has a single element, i.e., the unknown true data-generating distribution $p_\varepsilon(x, y)$, the inner maximization problem in Theorem 4.3.1 degenerates.

In solving (4.4), since $p_\varepsilon(x, y)$ is unknown, we resort to finding a tight upper bound of the inner maximization problem that can be evaluated without knowing the true joint distribution p_ε. Then in the outer minimization problem, we can instead optimize the surrogate upper bound and obtain the minimal loss (i.e., the distribution divergence) and the corresponding $\hat{\theta}$. The hope is that $\hat{\theta}$ is close to θ^*, the global optimal of the minimax problem (4.4). In the following, we present a tight upper bound that incorporates the testing phase distribution shift constraint.

Theorem 4.4.1. *Consider three joint distributions $\varepsilon, \phi, \theta \in (\mathcal{X} \times \mathcal{Y}, \mathcal{F})$ defined on a compact measure space. Suppose the following conditions are satisfied:*

- *The marginal probabilities* $p_\phi(x) = p_\theta(x), \forall x \in \mathcal{X}$,
- $D_{KL}[p_\varepsilon(x,y)||p_\phi(x,y)] \leq \alpha$,
- *there exists a constant* $\delta > 0$ *such that* $E_y(\delta) := \{\zeta | D_{KL}[p_y^{(\zeta)}||p_y^{(\phi)}] \leq \delta\}$ *is convex.*

then we have:

$$\begin{aligned} D_{KL}[p_\varepsilon(x,y)||p_\theta(x,y)] \leq I_\phi(X;Y) - I_\theta(X;Y) + \alpha + \|p_y^{(\varepsilon)} - p_y^{(\phi)}\|\sqrt{V(\log p_y^{(\phi)})} \\ + \|p_x^{(\varepsilon)} - p_x^{(\phi)}\|\sqrt{V(H_{y|X}^{(\phi)})} + \sqrt{E_{\varepsilon,x}\left[\|p_{y|X}^{(\varepsilon)} - p_{y|X}^{(\phi)}\|^2\right] E_{\varepsilon,x}\left[V(\log p_{y|X}^{(\phi)})\right]} \\ + D_{KL}[p_y^{(\phi)}||p_y^{(\theta)}], \end{aligned}$$

where $H_\theta(Y|X)$ *is the Shannon conditional entropy of* $p_\theta(x,y)$*;* $p_y^{(\theta)}$ *denotes the vector form of the marginal probability of* $p_\theta(y)$ *while* $p_{y|x}^{(\theta)}$ *denotes a* $|Y|$ *dimensional conditional probability vector given a realization of* $x \in \mathcal{X}$*; The expectation* $E_{\theta,x}[\,\cdot\,]$ *is taken with respect to the marginal probability* $p_\theta(x)$*;* $V(a) := \sum_{\mathrm{i}=1}^{|a|}(a_\mathrm{i} - \bar{a})^2$ *with* $\bar{a} := \frac{1}{|a|}\sum_{\mathrm{i}=1}^{|a|} a_\mathrm{i}$*, which resembles the equally weighted sample variance; Finally,* $H_{y|x}^{(\theta)}$ *is a* $|\mathcal{X}|$*-dimensional vector whose* i*-th entry is* $[H_{y|x}^{(\theta)}]_\mathrm{i} := -\sum_y p_\theta(y|x_\mathrm{i}) \log p_\theta(y|x_\mathrm{i}), \forall \mathrm{i} \in [|\mathcal{X}|]$.

Proof. Applying Theorem 4.3.1, assuming $m(\varepsilon,\theta) > m(\theta,\varepsilon)$, we have:

$$D_{KL}[p_\varepsilon(x,y)||p_\theta(x,y)] \leq I_\varepsilon(X;Y) - I_\theta(X;Y) + D_{KL}[p_\varepsilon(x)||p_\phi(x)] + D_{KL}[p_\varepsilon(y)||p_\theta(y)], \quad (4.5)$$

where $p_\phi(x) = p_\theta(x)$ and $p_\theta(y) = \sum_x p_\theta(y|x)p_\phi(x)$ are because of the assumptions in the statement. Then by the variational form of the mutual information, we rewrite:

$$\begin{aligned} I_\varepsilon(X;Y) - I_\theta(X;Y) &= H_\varepsilon(Y) - H_\theta(Y) - H_\varepsilon(Y|X) + H_\theta(Y|X) \\ &= H_\varepsilon(Y) - H_\phi(Y) + H_\phi(Y) - H_\theta(Y) \\ &\quad - H_\varepsilon(Y|X) + H_\phi(Y|X) - H_\phi(Y|X) + H_\theta(Y|X), \\ &= I_\phi(X;Y) - I_\theta(X;Y) + H_\varepsilon(Y) - H_\phi(Y) - H_\varepsilon(Y|X) + H_\phi(Y|X) \end{aligned}$$

where in the last equality we include the condition entropy of $p_\phi(x, y)$ to the equation. Then we find upper bounds for the first pair $H_\varepsilon(Y) - H_\phi(Y)$ and the second pair $H_\varepsilon(Y|X) - H_\phi(Y|X)$ separately. For the first pair, we have:

$$\begin{aligned} H_\varepsilon(Y) - H_\phi(Y) &= \sum_y p_\varepsilon(y) \log \frac{1}{p_\varepsilon(y)} - p_\phi(y) \log \frac{1}{p_\phi(y)} \\ &= \sum_y p_\varepsilon(y) \log \frac{1}{p_\varepsilon(y)} \frac{p_\phi(y)}{p_\phi(y)} - p_\phi(y) \log \frac{1}{p_\phi(y)} \\ &= -D_{KL}[p_\varepsilon(y)||p_\phi(y)] + \sum_y [p_\varepsilon(y) - p_\phi(y)] \left[\log \frac{1}{p_\phi(y)} - c \right]. \end{aligned} \tag{4.6}$$

Note that the constant c in the last line of the above equation can be any number independent of y. We simply pick c as the sample mean when treating $-\log p_\phi(y)$ as a vector of samples, that is:

$$\bar{c} := \frac{1}{|Y|} \sum_{\mathrm{i}=1}^{|Y|} \log \frac{1}{p_\phi(y_\mathrm{i})}. \tag{4.7}$$

Substitute (4.7) into (4.6), we arrive at:

$$H_\varepsilon(Y) - H_\phi(Y) = -D_{KL}[p_\varepsilon(y)||p_\phi(y)] + \sum_y [p_\varepsilon(y) - p_\phi(y)] \left[\log \frac{1}{p_\phi(y)} - \bar{c} \right].$$

Then for the last term in the above equation, applying Cauchy-Schwarz inequality to it, we obtain:

$$H_\varepsilon(Y) - H_\phi(Y) = -D_{KL}[p_\varepsilon(y)||p_\phi(y)] + \|p_y^{(\varepsilon)} - p_y^{(\phi)}\| \sqrt{V(\log p_y^{(\phi)})}, \tag{4.8}$$

where $p_y^{(\phi)}$ is the vector form of the marginal probability $p_\phi(y)$ and $V(\cdot)$ the sample variance as defined in the statement of Theorem 4.4.1. Then by assumption, the set $E_y(\delta)$ is a convex set for some $\delta > 0$, applying the Pythagorean theorem [65], we have:

$$D_{KL}[p_y^{(\varepsilon)}||p_y^{(\theta)}] - D_{KL}[p_y^{(\varepsilon)}||p_y^{(\phi)}] \geq D_{KL}[p_y^{(\phi)}||p_y^{(\theta)}]. \tag{4.9}$$

On the other hand, for the second term $-H_\varepsilon(Y|X) + H_\phi(Y|X)$, we have:

$$\begin{aligned}
-H_\varepsilon(Y|X) + H_\phi(Y|X) &= \sum_{x,y} p_\varepsilon(x)p_\varepsilon(y|x)\log p_\varepsilon(y|x) - p_\phi(x)p_\phi(y|x)\log p_\phi(y|x) \\
&= \sum_{x,y} p_\varepsilon(x)p_\varepsilon(y|x)\log p_\varepsilon(y|x) \\
&\quad - \sum_{x,y} \left[p_\phi(x) - p_\varepsilon(x) + p_\varepsilon(x)\right] p_\phi(y|x)\log p_\phi(y|x) \\
&= \sum_x p_\varepsilon(x)\left[\sum_y p_\varepsilon(y|x)\log p_\varepsilon(y|x) - p_\phi(y|x)\log p_\phi(y|x)\right] \\
&\quad + \sum_x \left[p_\varepsilon(x) - p_\phi(x)\right]\left[\sum_y p_\phi(y|x)\log p_\phi(y|x) - \bar{b}(x)\right] \\
&\leq \sum_x p_\varepsilon(x)\left[\sum_y p_\varepsilon(y|x)\log p_\varepsilon(y|x) - p_\phi(y|x)\log p_\phi(y|x)\right] \\
&\quad + \|p_x^{(\varepsilon)} - p_x^{(\phi)}\|\sqrt{V(H_{y|X}^{(\phi)})},
\end{aligned} \tag{4.10}$$

where the constant $\bar{b}(x)$, conditioned on $x \in \mathcal{X}$, introduced for the same reason as in (4.6), is defined as:

$$\bar{b}(x) := \frac{1}{|\mathcal{Y}|}\sum_y p_\phi(y|x)\log p_\phi(y|x).$$

Then, for the first term in the last inequality of (4.10), we have:

$$\begin{aligned}
&\sum_x p_\varepsilon(x)\left[\sum_y p_\varepsilon(y|x)\log p_\varepsilon(y|x) - p_\phi(y|x)\log p_\phi(y|x)\right] \\
=&\sum_x p_\varepsilon(x)\left[\sum_y p_\varepsilon(y|x)\log p_\varepsilon(y|x)\frac{p_\phi(y|x)}{p_\phi(y|x)} - \sum_y p_\phi(y|x)\log p_\phi(y|x)\right] \\
=&E_{x,\varepsilon}\left\{D_{KL}\left[p_\varepsilon(y|x)||p_\phi(y|x)\right]\right\} + \sum_x p_\varepsilon(x)\left\{\sum_y \left[p_\varepsilon(y|x) - p_\phi(y|x)\right]\left[\log p_\phi(y|x) - \bar{k}(x)\right]\right\} \\
\leq&E_{x,\varepsilon}\left\{D_{KL}\left[p_\varepsilon(y|x)||p_\phi(y|x)\right]\right\} + E_{x,\varepsilon}\left[\|p_{y|X}^{(\varepsilon)} - p_{y|X}^{(\phi)}\|\sqrt{V(\log p_{y|X}^{(\phi)})}\right] \\
\leq&E_{x,\varepsilon}\left\{D_{KL}\left[p_\varepsilon(y|x)||p_\phi(y|x)\right]\right\} + \sqrt{E_{x,\varepsilon}\left[\|p_{y|X}^{(\varepsilon)} - p_{y|X}^{(\phi)}\|^2\right]E_{x,\varepsilon}\left[V(\log p_{y|X}^{(\phi)})\right]},
\end{aligned} \tag{4.11}$$

where we apply Cauchy-Schwarz inequality twice for the last two inequalities; and the constant $\bar{k}(x)$, conditioned on $x \in \mathcal{X}$, is defined as:

$$\bar{k}(x) := \frac{1}{|Y|} \sum_{y} \log p_\phi(y|x).$$

Finally, for the first term in the last inequality of (4.10), we can exploit the assumption and obtain an upper bound that includes α:

$$D_{KL}[p_\varepsilon(x,y)||p_\phi(x,y)] = D_{KL}[p_\varepsilon(x)||p_\phi(x)] + E_{x,\varepsilon}\left\{D_{KL}\left[p_\varepsilon(y|x)||p_\phi(y|x)\right]\right\} \leq \alpha. \tag{4.12}$$

Putting all the pieces together, continuing from (4.5), we get:

$$\begin{aligned}
&D_{KL}[p_\varepsilon(x,y)||p_\theta(x,y)] \\
\leq &H_\varepsilon(Y) - H_\phi(Y) - H_\varepsilon(Y|X) + H_\phi(Y|X) + D_{KL}[p_\varepsilon(x)||p_\phi(x)] + D_{KL}[p_\varepsilon(y)||p_\theta(y)] \\
&\qquad + I_\phi(X;Y) - I_\theta(X;Y) \\
\leq &\|p_y^{(\varepsilon)} - p_y^{(\phi)}\|\sqrt{V(\log p_y^{(\phi)})} - H_\varepsilon(Y|X) + H_\phi(Y|X) + D_{KL}[p_\varepsilon(x)||p_\phi(x)] \\
&\quad + I_\phi(X;Y) - I_\theta(X;Y) + D_{KL}[p_\varepsilon(y)||p_\theta(y)] - D_{KL}[p_\varepsilon(y)||p_\phi(y)] \\
\leq &\|p_y^{(\varepsilon)} - p_y^{(\phi)}\|\sqrt{V(\log p_y^{(\phi)})} - H_\varepsilon(Y|X) + H_\phi(Y|X) + D_{KL}[p_\varepsilon(x)||p_\phi(x)] \\
&\qquad + I_\phi(X;Y) - I_\theta(X;Y) + D_{KL}[p_\phi(y)||p_\theta(y)] \\
\leq &\|p_y^{(\varepsilon)} - p_y^{(\phi)}\|\sqrt{V(\log p_y^{(\phi)})} + \|p_x^{(\varepsilon)} - p_x^{(\phi)}\|\sqrt{V(H_{y|X}^{(\phi)})} + D_{KL}[p_\varepsilon(x)||p_\phi(x)] \\
&+ E_{x,\varepsilon}\left\{D_{KL}\left[p_\varepsilon(y|x)||p_\phi(y|x)\right]\right\} + \sqrt{E_{x,\varepsilon}\left[\|p_{y|X}^{(\varepsilon)} - p_{y|X}^{(\phi)}\|^2\right] E_{x,\varepsilon}\left[V(\log p_{y|X}^{(\phi)})\right]} \\
&\qquad + I_\phi(X;Y) - I_\theta(X;Y) + D_{KL}[p_\phi(y)||p_\theta(y)] \\
\leq &\|p_y^{(\varepsilon)} - p_y^{(\phi)}\|\sqrt{V(\log p_y^{(\phi)})} + \|p_x^{(\varepsilon)} - p_x^{(\phi)}\|\sqrt{V(H_{y|X}^{(\phi)})} + \alpha \\
&+ \sqrt{E_{x,\varepsilon}\left[\|p_{y|X}^{(\varepsilon)} - p_{y|X}^{(\phi)}\|^2\right] E_{x,\varepsilon}\left[V(\log p_{y|X}^{(\phi)})\right]} + I_\phi(X;Y) - I_\theta(X;Y) + D_{KL}[p_\phi(y)||p_\theta(y)],
\end{aligned}$$

where the last four inequalities follow (4.8), (4.9), (4.10), (4.11) and (4.12) sequentially. □

While the upper bound in Theorem 4.4.1 still depends on the unknown distribution ε, but observe that the dependency is related to the 2-norm on the difference of probability

measures between the true ε and its estimate ϕ from finite sampling. Since any probability density is bounded between $[0, 1]$, by elementary results in learning theory, with enough number of samples m, the terms with 2-norm diminish to zero at a rate $\mathcal{O}(1/\sqrt{m})$ with high probability [14], [15], [88]. In this case, the upper bound reduces to:

$$D_{KL}[p_\varepsilon(x, y)||p_\theta(x, y)] \leq \alpha + I_\phi(X; Y) - I_\theta(X; Y).$$

The results imply that if the learner have enough samples about the true distribution, even when the learning model completely fits the training distribution $p_\phi(x, y)$, i.e., when $I_\theta(X; Y) = I_\phi(X; Y)$, the generalization error (distribution divergence) always reaches α eventually. In other words, if α is within an acceptable level to a learner, then there is no need to worry about overfitting once there is enough data to give accurate estimate of the data-generating distribution. However, this insight is based on the premise that enough samples about the true distribution are available to the learner, which is infeasible in general learning tasks.

Based on Theorem 4.4.1, we can further demonstrate the asymptotic sample complexity in standard learning settings under some additional assumptions commonly adopted in literature [15], [16]. The first tool to show this is an elementary concentration inequality in learning theory, which is summarized as follows [88]: Given m samples of a random vector X with dimension n, denoted the collection of samples as S, and define a vector function $\phi(x) : \mathbb{R}^n \mapsto \mathbb{R}^d$. The true mean is $E_x[\phi(x)]$ while the estimated one from S is denoted as $\phi_S = (1/m)\sum_{\mathrm{i}=1}^{m} \phi(x_\mathrm{i})$. If for any $x \in \mathcal{X}$, there exists a positive constant $C := \sup_{x\in\mathcal{X}}\|\phi(x)\|$. Then with probability of least $1 - \delta$, the estimation error for $\phi(x)$ satisfies:

$$\|\phi_S - E_x[\phi(X)]\| \leq \frac{C}{\sqrt{m}}\left(2 + \sqrt{2\log\frac{1}{\delta}}\right). \tag{4.13}$$

The above technique can be applied to the norms of the probability vectors since, for example, the probability vector $p_y^{(\varepsilon)}$ corresponds to the true distribution while $p_y^{(\phi)}$ is the empirical estimate from S and $\sup_{y\in\mathcal{Y}} p(y) = 1$. Therefore, we can replace $\|p_x^{(\varepsilon)} - p_x^{(\phi)}\|$, $\|p_y^{(\varepsilon)} - p_y^{(\phi)}\|$ and $\|p_{y|x}^{(\varepsilon)} - p_{y|x}^{(\phi)}\|$ with (4.13), to get an estimate of rate the concentration of the general-

ization error, and hence the asymptotic sample complexity. As for the sample variances involved in Theorem 4.4.1, for discrete settings the entropy is bounded from both sides $0 \leq H(Y|x) \leq \log |Y|$ whereas if smoothness assumptions are imposed on the (conditional) probability vectors, i.e., the ξ-infimality (Definition 2.4.2) holds for $p_y^{(\phi)}, p_{y|x}^{(\phi)}, \forall x \in \mathcal{X}, y \in \mathcal{Y}$, then $V(\log p_y^{(\phi)})$ and $V(\log p_{y|x}^{(\phi)})$ are bounded as the variance of a bounded variable $x \in [a, b]$ satisfies $V[X] \leq (b-a)^2/4$. Putting these results together, we have the following theorem.

Theorem 4.4.2. *Suppose the conditions in Theorem 4.4.1 are satisfied. In addition, if the followings hold:*

- $p_y^{(\varepsilon)}, p_y^{(\phi)}$ *are* ξ_y*-infimal.*
- $p_{y|x}^{(\varepsilon)}$ *is* $\xi_{y|x}$*-infimal* $\forall x \in \mathcal{X}$ *and* $\xi_{y|x}^* := \inf_{x \in \mathcal{X}} \xi_{y|x}$.

Then with probability of at least $1-\delta$ *and* m *samples drawn from the data-generating distribution* $p_\varepsilon(x, y)$*, the following inequality holds:*

$$D_{KL}[p_\varepsilon(x,y)||p_\theta(x,y)] \leq I_\phi(X;Y) - I_\theta(X;Y) + \alpha + \frac{\sqrt{\log \frac{|\mathcal{X}|}{\delta}}}{\sqrt{m}} \left[C(\xi_y) + C(\xi_{y|x}^*)\right].$$

where the constants $C(\xi_y), C(\xi_{y|x}^*)$ *are defined as:*

$$C(\xi_y) := \frac{\sqrt{C^*}}{2} \left[\log \frac{|\mathcal{Y}|}{\xi_y(1-\xi_y)}\right], \quad C(\xi_{y|x}^*) := -\frac{C^*}{4} \log \xi_{y|x}^*(1-\xi_{y|x}^*),$$

with $C^* \geq \left(2 + \sqrt{2 \log \frac{|\mathcal{X}|+2}{\delta}}\right)^2 / \log \frac{|\mathcal{X}|}{\delta}$ *is some small constant.*

Proof. By Theorem 4.4.1, and (4.13), adjusting the confidence to $\frac{\delta}{|\mathcal{X}|+2}$ due to the union bound on each term $\|p_y^{(\varepsilon)} - p_y^{(\phi)}\|$, $\|p_x^{(\varepsilon)} - p_x^{(\phi)}\|$ and $\|p_{y|x}^{(\varepsilon)} - p_{y|x}^{(\phi)}\|$ for each given $x \in \mathcal{X}$. Then with probability of at least $1-\delta$, we have:

$$\|p_y^{(\varepsilon)} - p_y^{(\phi)}\| \leq \frac{2 + \sqrt{2 \log \frac{|\mathcal{X}|+2}{\delta}}}{\sqrt{m}} \leq \sqrt{\frac{C^* \log \frac{|\mathcal{X}|}{\delta}}{m}}, \tag{4.14}$$

where the small constant C^* is defined as in the statement of the theorem. We can derive similar results for the other $|\mathcal{X}|+1$ norm of difference of probability vectors similarly. On the

other hand, for the variances, by assumption $p_y^{(\phi)}$ is ξ_y-infimal, which equivalently expressed as $\log(1-\xi_y) \geq \log p_\phi(y) \geq \log \xi_y, \forall y \in \mathcal{Y}$. Hence, we have:

$$V(\log p_y^{(\phi)}) \leq \frac{[\log \varepsilon_y(1-\varepsilon_y)]^2}{4}. \tag{4.15}$$

Note that we can derive similar result for $V(\log p_{y|x}^{\phi}), \forall x \in \mathcal{X}$. Lastly, for $V(H_{Y|x}^{\phi}), \forall x \in \mathcal{X}$, because $H_{Y|x}^{\phi}$ is the entropy function and we assume discrete $\mathcal{Y}$, hence each $0 \leq H_{Y|x}^{\phi} \leq \log|\mathcal{Y}|$, and we have:

$$Var(H_{Y|x}^{\phi}) \leq \frac{(\log|\mathcal{Y}|)^2}{4}. \tag{4.16}$$

Substitute (4.14)for norm of the difference of probability vectors, (4.15) and (4.16) for the variance function, and the fact that expectation of a constant gives the constant itself, we complete the proof. □

Theorem 4.4.2 implies that the asymptotic sample complexity for a confidence level δ reduces at a rate of $\mathcal{O}(\frac{1}{\sqrt{m}})$ with m denotes the number of samples needed. This rate improves over that in literature, which is $\mathcal{O}(\sqrt{\frac{\log m}{m}})$ [15]. However, the bound depends on the dimension of the observations as the asymptotic dependence is $\sqrt{\log|\mathcal{X}|}$, which is undesirable given the scale and amount of data in modern learning problems of practical interests. In addition, due to the additional assumptions of the smoothness of the marginal and conditional probability vectors $(\xi_y, \xi_{y|x}^*)$, the bound in Theorem 4.4.2, as in literature, is sensitive to the "outliers". While this weakness can be addressed through data-preprocessing, it is of the fundamental interests to incorporate these rare events into a unified theoretic framework. We leave this direction for future exploration.

4.4.2 Connection to Learning through IB Methods

In this part, we connect the proposed framework to the IB methods that have recently been applied to DNN through variational inference and therefore allows efficient estimation [2]. Consider a supervised-learning scenario, where the random variable X denotes the observation; Y the target and Z the latent features. Then recall in IB methods, the Markov chain: $Y - X - Z$ holds and one attempts to find a set of encoders conditional probability

$p(z|x)$ from training distribution such that Z attains maximal relevance measured in $I(Y;Z)$ while minimal complexity $I(X;Z)$ is achieved for a given trade-off constraint controlled by a Lagrangian multiplier $\beta > 0$. Due to the Markov chain, by strong data-processing inequality (SDPI [85], [86], [89]): $\eta_F I(Z;X) \geq I(Z;Y)$, where η_F denotes the SDPI coefficient with $0 < \eta_F \leq 1$. However, prior works often treated the target variable Y as the prediction $\hat{Y}$ for convenience of expression while the full Markov chain involved is $Y - X - Z - \hat{Y}$, as identified recently [90]. Focusing on the Markov chain that involved in the prediction process at the testing phase, that is $X - Z - \hat{Y}$, by SDPI, we have $I(\hat{Y};X) \leq \eta_B I(Z;X)$, where η_B denotes the SDPI coefficient where $0 < \eta_B \leq 1$. We can incorporate this relation to Theorem 4.4.1, which gives the next result.

Theorem 4.4.3. *For the prediction Markov chain $X - Z - \hat{Y}$, then we have:*

$$\begin{aligned} D_{KL}[p_\varepsilon(x,y)||p_\theta(x,\hat{y})] \leq I_\phi(X;\hat{Y}) + \eta_B I_\theta(Z;X) - I_\theta(X;\hat{Y}) + \alpha \\ + \|p_y^{(\varepsilon)} - p_y^{(\phi)}\| \sqrt{V(\log p_y^{(\phi)})} + \|p_x^{(\varepsilon)} - p_x^{(\phi)}\| \sqrt{V(H_{y|X}^{(\phi)})} \\ + D_{KL}[p_y^{(\phi)}||p_y^{(\theta)}] + \sqrt{E_{\varepsilon,x}\left[\|p_{y|X}^{(\varepsilon)} - p_{y|X}^{(\phi)}\|^2\right] E_{\varepsilon,x}\left[V(\log p_{y|X}^{(\phi)})\right]}, \end{aligned}$$

where η_{KL} denotes the strong data-processing coefficient with $0 < \eta_B \leq 1$ and is defined as:

$$\eta_B := \sup_{p_\theta(y,z)\in\Theta} \frac{I_\theta(X;\hat{Y})}{I_\theta(Z;X)}.$$

Proof. Without loss of generality, assume $m(\varepsilon,\theta) > m(\theta,\varepsilon)$. By the Markov chain $X - Z - \hat{Y}$, we adopt the strong data-processing inequality:

$$\eta_B I_\theta(Z;X) \geq I_\theta(X;\hat{Y}).$$

This implies that $\eta_B I(Z;X) - I(\hat{Y};X) \geq 0$, which serves as a regularized upper bound and aligns with the common practice in recent research. Substitute the above into the inequality of Theorem 4.4.1, we complete the proof. □

Recall that in the proposed framework, Theorem 4.4.3 serves as a surrogate upper bound of the inner maximization problem, when minimizing this upper bound with respect to the learning model θ, the equivalent loss reduces to:

$$\mathcal{L}_\theta := \eta_B I_\theta(Z;X) - I_\theta(X;\hat{Y}),$$

which reveals the predictive IB Lagrangian as adopted in recent works [87]. Note that the term $D_{KL}[p_y^{(\phi)}||p_y^{(\theta)}] = 0$ under the Markov chain $Y - X - Z - \hat{Y}$ which can be shown easily along with the Bayes' rule. On the other hand, although in Theorem 4.4.3 we assume $m(\varepsilon,\theta) > m(\theta,\varepsilon)$, if the opposite holds instead, then the corresponding upper bound is the same. To see this, denote the equivalent loss function of the later case as $\bar{\mathcal{L}}_\theta$:

$$\bar{\mathcal{L}}_\theta := I_\theta(X;\hat{Y}) - I_\phi(X;Y) \leq \eta_B I_\theta(X;Z) - I_\theta(X;\hat{Y}),$$

where the first inequality follows the predictive Markov chain $X - Z - \hat{Y}$ and the second inequality follows from data-processing inequality, that is, the learned model cannot acquire more mutual information than the training data has. This result demonstrates that the IB methods is a surrogate loss upper bound in our framework, we evaluate it in a synthetic dataset in Chapter 4.5.2.

Remarkably, due to the Markov chain involved, this result extends to a broader class of learning model with the encoder-decoder architecture [91].

As for the SDPI coefficient η_B with the Bayes' decoders obtained through IB, there are existing results in finding η_B numerically [73], [74], [90]. These works study the second order statistics of IB methods from the viewpoint of bifurcation theory, probability theory respectively and more recently [76] that linked the SDPI coefficient of IB to the maximum correlation coefficient and generalizes them to continuous settings through perturbation analysis. Among which, [74] provides the simplest form that suits our purpose in applying Theorem 4.4.3. The result is summarized as follows:

Assume the joint probability $p(x, y)$ is known. Suppose the Markov chain $Y - X - Z$ holds and the conditional probability $p(z|x)$ satisfies:

$$p(z|x) := \arg\min_{p(z|x)\in\Omega} \quad \eta_{KL} I(X;Z) - I(Y;Z).$$

Define $\lambda_2(z), \forall z \in \mathcal{Z}$ as the second largest singular value of the matrix: $Q_z := \Lambda_{y|z}^{-\frac{1}{2}} P_{Y|X} \Lambda_{x|z}^{\frac{1}{2}}$. Then we have $\eta_{KL} := \max_{z\in\mathcal{Z}} \lambda_2(z)$. Note that $Q_z^T Q_z$ is the Hessian matrix, conditioned on $z \in \mathcal{Z}$, of the IB Lagrangian functional with $p(z|x)$ as the variables.

4.4.3 Tightening Existing Upper Bounds

In this part, we aim at applying Theorem 4.3.1 on the existing mutual information-based generalization error upper bounds. But before presenting the results, we first establish the common ground for comparison. In literature, the mutual information is measured between the input and output of a learning model. The input, denoted as $S_m^{(\mu)}$, is the dataset of m samples of observation-label pairs (X, Y) drawn i.i.d from a certain data-generating process μ, which gives

$$S_m^{(\mu)} := \{z_{\mathrm{i}}\}_{\mathrm{i}=1}^m, \quad z_{\mathrm{i}} = (x_{\mathrm{i}}, y_{\mathrm{i}}) \sim \mu \in \mathcal{X} \times \mathcal{Y}.$$

On the other hand, the output refers to the hypothesis $\theta \in \Theta(A)$, generated with an algorithm A. The hypothesis, or equivalently, the learned model θ is produced through a Markov kernel $p(\Theta|S_m^{(\mu)})$. Then, given a σ-sub-Gaussian measurable loss function $l(s, \theta)$, the population risk with an unknown data-generating distribution μ is $L_\mu(\theta) := E_\mu[L(S_\infty, \theta)]$ while the empirical risk $L_S^N(\theta) := (1/N)\sum_{\mathrm{i}=1}^N l(z_{\mathrm{i}}, \theta), \forall z_{\mathrm{i}} \in S_N$. Note that we can think of the population risk, whose evaluation still needs a realization of a hypothesis $\theta \in \Theta$. In this setup, the generalization error is defined as:

$$\mathrm{gen}(A, S_m, \Theta) := \left| E_\theta \left[L_\mu(\Theta) - L_S^m(\Theta) \right] \right|, \tag{4.17}$$

Then by the Donsker-Varadhan's representation of the KL divergence as shown in literature [18]:

$$gen(A, S_m, \Theta) \leq \sqrt{2\sigma^2 I(S;\Theta)}.$$

Note that the outer expectation provides the hypothesis class for the inner expectations and hence they can be evaluated accordingly. Mathematically, the corresponding mutual information is:

$$I(S,\Theta) := \sum_{s,\theta} p(s)p(\theta|s) \log \frac{p(\theta|s)p(s)}{p(\theta)} = \sum_{x,y} p_\mu(x)p_\theta(y|x) \log \frac{p_\theta(y|x)}{p_\theta(y)}.$$

Observe that to make the expression to be a mutual information, the denominator must satisfy $p_\theta(y) = \sum_x p_\mu(x)p_\theta(y|x)$, or equivalently seeing it as a divergence metric, the KL divergence between joint measure $p_\mu(x)p_\theta(y|x)$ to its product measure $p_\mu(x)p_\theta(y)$. To relax the above-mentioned restrictions, we propose another form of the generalization error gap from a reverse direction compared to previous works.

Our motivation is the insights learned through Theorem 4.4.1, where it implies that overfitting is not an issue if there are enough number of samples, or the distribution mismatch is within an acceptable threshold. Therefore, different from existing formulation, we consider the case where the learning model also depends on the data samples, or equivalently, the training distribution. In details, define a σ-sub-Gaussian measurable function $\lambda f(S_m, \Theta(S_m, A))$ which is viewed as the sample average of the single instance loss function $l(z,\theta)$ for some $\lambda \in \mathbb{R}$. The hypothesis space $\Theta(S_m, A)$ depends on the available data, or equivalently the training distribution and output a hypothesis θ through an algorithm A. Then we consider the following difference:

$$E_\varepsilon[\lambda f(S_\infty, \Theta(S_\infty))] - E_\phi[\lambda f(S_m, \Theta(S_m))],$$

where if ε is the unknown data-generating distribution, then by weak law of large number the first term of the above is equivalent to:

$$E_\varepsilon[\lambda f(S_\infty, \Theta(S_\infty))] = \lim_{m\to\infty} \frac{1}{m} \sum_{\mathrm{i}=1}^{m} l(z_\mathrm{i}, \theta_m). \tag{4.18}$$

On the other hand, the second term is the empirical risk, where ϕ is viewed as the empirical estimate of ε from the available finite samples:

$$E_{\phi}[\lambda f(S_m, \theta_m)] = \frac{1}{m}\sum_{\mathrm{i}=1}^{m} l(z_{\mathrm{i}}, \theta_m) \tag{4.19}$$

Note that in this set up, a realization of the learning model θ_m depends on the available samples. Another observation is that in (4.19), the m samples are drawn from the unknown distribution ε which takes the intrinsic sampling distribution mismatch into account, since $S_m = \{z_{\mathrm{i}}\}_{\mathrm{i}=1}^{m} \sim \varepsilon$ but the learning model θ_m is fitting toward the empirical distribution ϕ, which is estimated from S_m. Following this formulation, we define the generalization error gap as:

$$\mathrm{gen}(A, S_m, \Theta) := \left| E_{\theta}\left[E_{\varepsilon}[L(S_{\infty}, \Theta_{\infty})] - E_{\phi}[L(S_m, \Theta_m)] \right] \right|, \tag{4.20}$$

where we rewrite $L(S_m, \Theta_m) := \lambda f(S_m, \Theta_m)$. Compared to previous form, the main difference of (4.20) is that: we include the training distribution, estimated from the finite samples and therefore the model intrinsically deviates from the true data-generating distribution if overfitted. Then due to the σ-sub-Gaussian of the loss function, the KL divergence in the upper bound can be expressed similarly to that derived in the input-output mutual information. But now, the upper bound that corresponds to the proposed form is:

$$\mathrm{gen}(A, S_m, \Theta) \leq \sqrt{2\sigma^2 D_{KL}[p_{\varepsilon}(x, y)||p_{\theta}(y|x)p_{\phi}(x)]}.$$

Then clearly, we can apply Theorem 4.3.1 to the KL divergence. In comparing to existing bounds, it turns out that this form can result in a better estimate of the generalization error gap than that obtained from (4.20).

Theorem 4.4.4. *Let ε, ϕ be two Borel measurable functions defined over a compact support $B := \{\mu | \mu(z) > 0, \forall z = (x, y) \sim \mu : (\mathcal{X}, \mathcal{Y}) \mapsto \mathbb{R}\}$ and $gen(A, S_m, \Theta)$ is defined as in (4.20). Suppose the loss function $L(S, \Theta)$ of a finite dataset S and the hypothesis space $\Theta(A)$ is σ-sub-Gaussian, then we have:*

$$gen(A, S_m, \Theta) \leq \sqrt{2\sigma^2 \max\{m(\varepsilon, \theta), m(\theta, \varepsilon)\}},$$

where the function of two joint measures $\varepsilon, \phi \in B$ is defined as:

$$M(\varepsilon, \theta) := I_\varepsilon(X;Y) - I_\theta(X;Y) + D_{KL}[p_\varepsilon(x)||p_\phi(x)] + D_{KL}[p_\varepsilon(y)||p_\theta(y)],$$

Proof. The proof starts from the similar steps of of [17], that is, the Donsker-Varadhan's representation of the KL divergence:

$$E_{\varepsilon,\theta}[L(S_\infty, \theta_\infty)] - E_{\phi,\theta}[L(S_m, \Theta_m)] \leq \sqrt{2\sigma^2 D_{KL}[p_\varepsilon(x,y)||p_\theta(x,y)]}.$$

Then we apply Theorem 4.3.1 to the KL divergence in the above inequality. Since the generalization error gap is defined as the absolute difference between the two expectations, by interchanging the order of ε and ϕ in the above inequality and we complete the proof. □

As a remark, in Theorem 4.4.4, if we substitute $p_\theta(x,y) = p_\varepsilon(x)p_\varepsilon(y)$ then we recover the results in [17], [18], so our results generalize the mutual information-based generalization error bounds to mismatch distributions beyond the product measures. In addition, as commonly assumed in this line of research, a divergence constraint is imposed on the feasible set of hypotheses, or the stability conditions [17], [18], [60]: $I_\varepsilon(X;Y) \leq \alpha$. If we generalize the product measure in the mutual information to the joint distribution of the training data $p_\phi(x,y)$, then the stability conditions is simply the KL divergence: $D_{KL}[p_\varepsilon(x,y)||p_\phi(x,y)] \leq \alpha$, which is the distribution mismatch constraint defined in our problem formulation 4.2.

4.4.4 Application on Existing Tightening Techniques

In this part, to demonstrate the benefit of the proposed methods, we apply our main results to various existing upper-bound tightening techniques for the input-output mutual information generalization error bounds. As discussed in chapter 1.2.4, [60] relaxes the sub-Gaussian assumption by treating the process of maximizing rate of decay of the deviation probability due to the application of the Chernoff inequality as a general inverse of a finite upper bound of the CGF of a random variable, assuming the existence of such function. This assumption is easily satisfied because of the Donsker-Varadhan's represen-

tation of the KL divergence. To better highlight the difference before and after applying our results, we summarize the results of [60] for self-contained purposes. Denote the CGF of a Borel-measurable function $f(x,y) \in B := \{\mu | \mu(x,y) > 0, \forall \mu \in \mathcal{X} \times \mathcal{Y}\}$ as $K(s) := \log E_\mu[\exp\{sf(x,y) - sE[f(X,Y)]\}]$ for $s \in (c_-, c_+), c_+ > 0, c_- < 0$, assume that there exists some functions $\Psi_+(s), \Psi_-(s)$ such that $K(s) \leq \Psi_+(s)$ for $s \in [0, c_+)$ while $K(s) \leq \Psi_-(-s)$ for $s \in (c_-, 0]$, and $\Psi_+(0) = \Psi_-(0) = \Psi_+(0) = \Psi_-(0) = 0$. Starting from the KL divergence between two joint measures $\mu, \nu \in B$, for $s \in [0, c_+)$ we have:

$$\begin{aligned} D_{KL}[p_\mu(x,y)||p_\nu(x,y)] &\geq sE_\mu[p_\mu(x,y)] - \log E_\nu[\exp\{sp_\nu(x,y)\}] \\ &\geq s\left[E_\mu[f(x,y)] - E_\nu[f(x,y)]\right] - \Psi_+(s), \end{aligned} \tag{4.21}$$

then by the definition of the generalized inverse [92], (4.21) can be written as:

$$E_\mu[f(x,y)] - E_\nu[f(x,y)] \leq \inf_{s \in [0, c_+)} \frac{D_{KL}[p_\mu(x,y)||p_\nu(x,y)] + \Psi_+(s)}{s}.$$

On the other hand, for the case where $s \in (c_-, 0]$, following similar steps (deferring the details to [60]), we get:

$$E_\mu[f(X,Y)] - E_\nu[f(X,Y)] \leq \inf_{s \in [0, -c_-)} \frac{D_{KL}[p_\mu(x,y)||p_\nu(x,y)] + \Psi_-(s)}{s} = \Psi^{*-1}(D_{KL}[\mu||\nu]).$$

Note that the exchange of the order of μ, ν is due to the range of $s \in (c_-, 0]$. This result generalizes the sub-Gaussian assumptions and hence includes a broader class of loss functions. Remarkably, this result recovers the σ-sub-Gaussian case as $\Psi_+^{*-1}(y) = \Psi_-^{*-1}(y) = \sqrt{2\sigma^2 y}$. In addition, observe that the result generalizes the mutual information upper bounds based on the choice of the two measures as $\mu = p(s,w), \nu = p(s)p(w)$, we can simply apply Theorem 4.3.1 on the KL divergence in 4.21 for two arbitrary joint measure, which results in the following generalization beyond sub-Gaussian satisfying loss functions:

Theorem 4.4.5. *Define the* $K(s) := \log E_\varepsilon[\exp\{s(f - E[f])\}]$ *the CGF of a measurable function* $f \in B$. *Suppose there exists some functions* $\Psi_+(s)$ *and* $\Psi_-(s)$ *with* $\Psi_+(0) = \Psi_-(0) =$

$\Psi_+(0) = \Psi_-(0) = 0$ *such that* $K(s) \leq \Psi_+(s), \forall s \in [0, c_+)$ *and* $K(s) \leq \Psi_-(-s), \forall s \in (c_-, 0]$ *are satisfied. Then for two measures* $\varepsilon, \phi \in B$*, we have:*

$$E_\varepsilon[f(x,y)] - E_\phi[f(x,y)] \leq \inf_{s \in [0, c_+)} \frac{M(\varepsilon, \phi) + \Psi_+(s)}{s},$$

and

$$E_\phi[f(x,y)] - E_\varepsilon[f(x,y)] \leq \inf_{s \in [0, -c_-)} \frac{M(\varepsilon, \phi) + \Psi_-(s)}{s}.$$

where $M(\mu, \nu) := \max\{m(\mu, \nu), m(\nu, \mu)\}$ *and:*

$$m(\mu, \nu) := I_\mu(X;Y) - I_\nu(X;Y) + D_{KL}[p_x^{(\mu)} || p_x^{(\nu)}] + D_{KL}[p_y^{(\mu)} || p_y^{(\nu)}].$$

Proof. By the Donsker-Varadhan's representation of the KL divergence, we have:

$$E_\varepsilon[sf(x,y)] - \log E_\phi[\exp\{sf(x,y)\}] \leq D_{KL}[\varepsilon || \phi], \quad \forall s \in \mathbb{R}.$$

Then by assumption, for $s \in [0, c_+)$, the CGF of $f(x,y)$ measured in ϕ is upper bounded by $\Psi_+(s)$, that is $K(s) := \log E_\phi[\exp\{s(f - E_\phi[f])\}] \leq \Psi_+(s)$, which gives:

$$\begin{aligned}
sE_\varepsilon[f(x,y)] - sE_\phi[f(x,y)] \leq & D_{KL}[p_\varepsilon(x,y) || p_\phi(x,y)] + \Psi_+(s) \\
\leq & D_{KL}[p_\varepsilon(x,y) || p_\phi(x,y)] + K(s) \\
\leq & \Big[I_\varepsilon(X;Y) - I_\phi(X;Y) + D_{KL}[p_x^{(\varepsilon)} || p_x^{(\phi)}] \\
& \qquad + D_{KL}[p_y^{(\varepsilon)} || p_y^{(\phi)}] \Big] + \Psi_+(s), \\
= & m(\varepsilon, \phi) + \Psi_+(s) \\
\leq & M(\varepsilon, \phi) + \Psi_+(s),
\end{aligned}$$

where we apply Theorem 4.3.1 to have the third inequality. Then optimizing the upper bound over s, followed by the definition of the generalized inversion [92], we complete the

first part of the theorem. On the other hand, for the case where $s \in (c_-, 0]$, we have similar results as shown in the following:

$$\begin{aligned} |s|\left(E_\phi[f(x,y)] - E_\varepsilon[f(x,y)]\right) &\leq D_{KL}[\varepsilon||\phi] + \Psi_-(|s|) \\ &\leq M(\varepsilon,\phi) + \Psi_-(|s|). \end{aligned}$$

Combining the two inequality we complete the proof. □

Another line of research that generalizes the mutual information-based bounds is by including a broader class of information-theoretic divergence, known as the α-information. In resemblance of the Pythagorean theorem [65] for the KL divergence, the Generalized Pythagorean theorem [63] is introduced along with the notion of α-convexity to generalize the convex set conditions as in the case for the Pythagorean theorem. Based on these results, we generalize Theorem 4.3.1 to the α-information-based divergence measures and hence show that our results apply to the upper-bound tightening techniques introduced in [63]. Similarly, before presenting Theorem 4.4.6, we summarize the results, mostly follows [58], [63] and refers the details therein. To begin with, this method is based on the Chernoff inequality:

$$P\left(\exp\{sf(x,y)\} \geq \frac{1}{\delta}\right) \leq E[\exp\{sf(x,y)\}]\delta, \tag{4.22}$$

where $f(x,y) \in B$ is some measurable function defined over a compact support. Then if there exists a non-negative random variable $U := sf(x,y)$ with $U \leq 1$, then (4.22) can equivalently rewritten as:

$$P\left(\exp\{sf(x,y)\} < \frac{1}{\delta}\right) \geq 1 - \delta.$$

This results can be adopted to establish learning-theoretic type of generalization error bounds, specifically, a deviation event happens with a probability that is dependent on to a confidence level. Consider the generalization error gap due to mismatch between the

unknown data-generating distribution ε and the sampling distribution ϕ. Since the learning model θ depends on the sample distribution ϕ:

$$\text{gen}(\phi, \theta) = E_\phi[L(S, \theta)] - E_\varepsilon[L(S, \theta)].$$

If in addition, the loss function $L(S, \theta)$ is σ-sub-Gaussian, then for all $\lambda \in \mathbb{R}$, we have:

$$E_\phi \left[\exp\left\{\lambda \left[E_\phi[L(S, \theta)] - E_\varepsilon[L(S, \theta)]\right]\right\}\right] \leq \exp\left\{\frac{\lambda^2\sigma^2}{2}\right\},$$

which can be rearranged to:

$$E_\phi \left[\exp\left\{\lambda \left[E_\phi\left[L(S, \theta)\right] - E_\varepsilon[L(S, \theta)]\right] - \frac{\lambda^2\sigma^2}{2}\right\}\right] \leq 1.$$

Note that the outer expectation is taken with respect to the sampling measure ϕ but we are interested in the case where the expectation is taken with respect to the true data-generating distribution ε. By "change of measure", we have:

$$\begin{aligned}
&E_\phi \left[\exp\left\{\lambda \left[E_\phi\left[L(S, \theta)\right] - E_\varepsilon[L(S, \theta)]\right] - \frac{\lambda^2\sigma^2}{2}\right\}\right] \\
&= \int d\varepsilon \frac{d\phi}{d\varepsilon} \exp\left\{\lambda \left[E_\phi\left[L(S, \theta)\right] - E_\varepsilon[L(S, \theta)]\right] - \frac{\lambda^2\sigma^2}{2}\right\} \\
&= E_\varepsilon \left[\exp\left\{\log \frac{d\phi}{d\varepsilon} + \lambda \left[E_\phi\left[L(S, \theta)\right] - E_\varepsilon[L(S, \theta)]\right] - \frac{\lambda^2\sigma^2}{2}\right\}\right],
\end{aligned}$$

where $d\phi/d\varepsilon$ denotes the Radon-Nikodym derivative [93]. Substituting the above into (4.22) results in an expression of the probability of the occurrence of the deviation event of interests:

$$P_\varepsilon \left(\exp\left\{\log \frac{d\phi}{d\varepsilon} + \lambda \left[E_\phi[L(S, \theta)] - E_\varepsilon[L(S, \theta)]\right] - \frac{\lambda^2\sigma^2}{2}\right\} \geq \frac{1}{\delta}\right) \leq \delta.$$

Equivalently, this implies with a probability at least $1 - \delta$, the following inequality holds:

$$\lambda \left\{E_\phi[L(S, \theta)] - E_\varepsilon[L(S, \theta)]\right\} \leq_{1-\delta} \log \frac{d\varepsilon}{d\phi} + \frac{\lambda^2\sigma^2}{2} + \log \frac{1}{\delta}, \tag{4.23}$$

where the subscript in $\leq_{1-\delta}$ denotes the confidence level of the inequality as denoted in literature [58], [63]. Next, we focus on the ratio $d\varepsilon/d\phi$. Recall the definition of α-information:

$$D_\alpha(\varepsilon||\phi) = \frac{1}{\alpha-1}\log\int d\varepsilon\left(\frac{d\varepsilon}{d\phi}\right)^{\alpha-1}.$$

Applying the Chernoff inequality to the logarithmic of the ratio of the measures: $\log d\phi/d\varepsilon$, with a constant $\alpha > 1$, we have:

$$\begin{aligned} P_\varepsilon\left(\log\frac{d\varepsilon}{d\phi}\geq t\right) &\leq E_\varepsilon\left[\exp\left\{(\alpha-1)\log\frac{d\varepsilon}{d\phi}\right\}\right]\cdot\exp\left\{-(\alpha-1)\,t\right\} \\ &= E_\varepsilon\left[\exp\left\{\log\left(\frac{d\varepsilon}{d\phi}\right)^{\alpha-1}\right\}\right]\cdot\exp\left\{-(\alpha-1)\,t\right\} \\ &= \exp\left\{(\alpha-1)\,D_\alpha(\varepsilon||\phi)-(\alpha-1)\,t\right\}. \end{aligned}$$

Then by letting $t = D_\alpha(\varepsilon||\phi) + \frac{1}{\alpha-1}\log\frac{1}{\delta}$, we conclude that with a probability of at least $1-\delta$:

$$\log\frac{d\varepsilon}{d\phi} \leq_{1-\delta} D_\alpha(\varepsilon||\phi) + \frac{1}{\alpha-1}\log\frac{1}{\delta}. \tag{4.24}$$

Substitute (4.23) into (4.23), which is done by the union bound for the event when both cases occur, we have for $\alpha > 1$, the following inequality holds:

$$\lambda\left[E_\phi[L(S,\theta)] - E_\varepsilon[L(S,\theta)]\right] \leq_{1-\delta^*} \frac{\lambda^2\sigma^2}{2} + D_\alpha(\varepsilon||\phi) + \left(\frac{\alpha}{\alpha-1}\right)\log\frac{2}{\delta^*}, \quad \forall\lambda\in\mathbb{R},$$

where we rewrite $\delta = \delta^*/2$, for the conciseness of expression. Since the above holds for all λ, and observe that it can be viewed as a quadratic programming problem with respect to λ, the corresponding discriminant must be negative, which results in:

$$E_\phi[L(S,\theta)] - E_\varepsilon[L(S,\theta)] \leq_{1-\delta^*} \sqrt{2\sigma^2\left[D_\alpha(\varepsilon||\phi) + \left(\frac{\alpha}{\alpha-1}\right)\log\frac{2}{\delta^*}\right]}, \tag{4.25}$$

where $\alpha > 1$. Observe that in the range of $\alpha > 1$, the coefficient pre-multiplied by the α-information $D_\alpha(\varepsilon||\phi)$ is positive, which allow us to apply the Generalized Pythagorean theorem [94] to (4.25). However, as mentioned in the reference work, the generalization

requires a notion known as the α-convexity which is defined in Now we are ready to present the next theorem.

Theorem 4.4.6. *Suppose a loss function $L(S,\theta)$ defined over a finite data samples and a learning model θ is σ-sub-Gaussian. For $\alpha \in (1,\infty)$, let ε, θ belong to an α-convex set of distributions $\mathcal{A}$ and ϕ be an arbitrary distribution. If $\theta := \arg\inf_{\mu\in\mathcal{A}} D_\alpha(\mu||\phi), \forall \mu \in \mathcal{A}$, then with probability $1-\delta$:*

$$E_\phi[L(S,\theta)] - E_\varepsilon[L(S,\theta)] \leq_{1-\delta} \sqrt{2\sigma^2\left[D_\alpha(\varepsilon||\phi) - D_\alpha(\theta||\phi) + \left(\frac{\alpha}{\alpha-1}\log\frac{2}{\delta}\right)\right]}.$$

Proof. The proof is based on the Generalized Pythagorean theorem [94], which states that for $\alpha \in (0,\infty)$:

$$D_\alpha(P||Q) \geq D_\alpha(P||P^*) + D_\alpha(P^*||Q),$$

where $P \in \mathcal{A}$ with $\mathcal{A}$ be a α-convex set and $P^* := \arg\inf_{P\in\mathcal{A}} D_\alpha(P||Q)$. In applying this result, we assign $P \leftarrow \varepsilon, Q \leftarrow \phi$ and $P^* \leftarrow \theta$. And since $(1,\infty) \subset (0,\infty)$, we can apply this to (4.25) and complete the proof. □

Remarkably, by the well-known property of the α-information in literature [94]:

$$\lim_{\alpha\to 1} D_\alpha(P||Q) = D_{KL}(P||Q),$$

letting $p_\phi(x,y) = p_\varepsilon(x)p_\varepsilon(y)$, then we have:

$$\begin{aligned} &D_{KL}[p_\varepsilon(x,y)||p_\varepsilon(x)p_\varepsilon(y)] - D_{KL}[p_\phi(x,y)||p_\varepsilon(x)p_\varepsilon(y)] \\ &\qquad = I_\varepsilon(X;Y) - I_\phi(X;Y) + D_{KL}[p_\varepsilon(x)||p_\phi(x)] + D_{KL}[p_\varepsilon(y)||p_\phi(y)], \end{aligned}$$

which recovers Theorem 4.3.1.

4.5 Numerical Results

In this part, we evaluate a variety of the generalization error upper bounds derived through the proposed technique on some synthetic datasets. We compare the results to the

input-output mutual information generalization upper bound that is recently introduced and actively studied. As shown in the Chapter 4.4, since the proposed technique applies to most existing tightening methods, for the compared bound, we will focus on the basic form [17], [18].

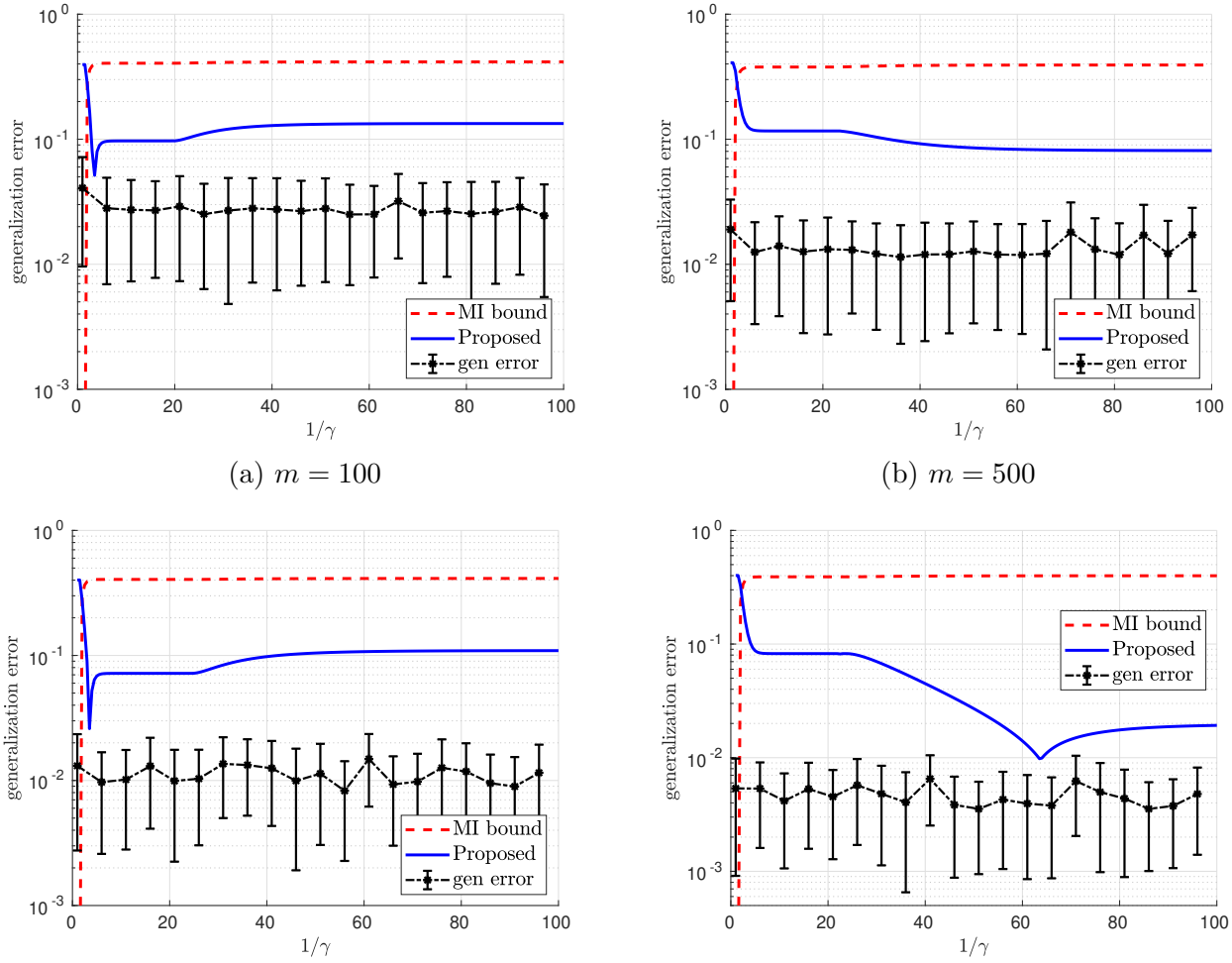
The simulation is divided into three parts. First, we consider standard learning task in supervised learning scenario where the goal is to bound the generalization error gap, that is, the absolute difference of the accuracy rates between training and testing phases. Second, we examine empirically the performance of IB methods in minimizing the maximum generalization error with the focus on comparing the optimal level of fitting between training and testing distributions. Lastly, we consider the most challenging, adversarial scenario where the goal is to find a proper level of fitting that minimize the distribution mismatch that is caused by both the sampling process and an adversary who has access to training data but is under some divergence constraints.

4.5.1 Standard Supervised Learning Task

In this part, we consider standard supervised learning task where there are two sets of data, where one is for training a learning model while the other is used to evaluate the performance of the learned model. To generate the two sets of data, we perform inverse transform sampling on the following synthetic joint probability.

$$P(Y|X) = \begin{bmatrix} 0.90 & 0.76 & 0.06 \\ 0.10 & 0.24 & 0.94 \end{bmatrix}, \quad P(X) = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}. \tag{4.26}$$

As for the learning model, we focus on the Bayes' decoder obtained through the IB objective with the training joint distribution estimated through counting the occurrence of the labels and observations. The benefit of this specific type of learning model is that the level of fitting the training distribution is reduced to a specific choice of the trade-off parameter γ.



(a) $m = 100$ (b) $m = 500$

(c) $m = 1000$ (d) $m = 5000$

Figure 4.1. Comparing generalization error gap upper bounds in a synthetic standard supervised learning tasks

For $\gamma \to 0$, this corresponds to overfitting case while $\gamma \to 1$ a complete uniformly randomized model. The Bayes' decoder is obtained as follows:

$$p_\theta(\hat{y}|x) = \sum_z p(\hat{y}|z)p(z|x) = \sum_z p(z|x)\frac{\sum_x p(z|x)p(x,y)}{\sum_x p(z|x)p(x)}.$$

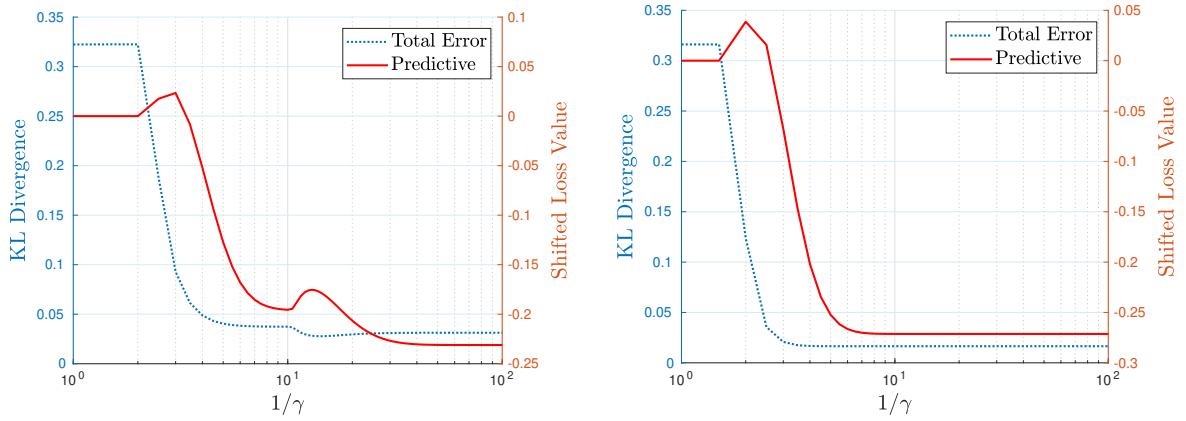
Note that $p_\theta(\hat{y}) = \sum_x p(x)p(\hat{y}|x) = p(y)$. The simulation starts with sampling the corresponding joint probability from (4.26) for m times, where $m \in \{100, 500, 1000, 5000\}$ as training dataset and each data is a pair of $(x_\mathrm{i}, y_\mathrm{i}), \mathrm{i} \in [m]$. Then we also sample an additional set of size 5000 as the testing data. As (4.26) is a simple example, 5000 samples is

a relatively large number, and we therefore assume the distribution mismatch in sampling testing data is negligible. Continuing the simulation, we run the IB algorithm with the joint probability $\hat{p}_\phi(x, y)$ comes from counting the training dataset. Then we can sweep through $\gamma \in (0.01, 1)$ where for each given γ, we restart the IB algorithm for 100 times and pick the encoder probability $p_\theta(z|x)$ with the highest $I(\hat{Y}; X)$ as the kernel to form the output hypothesis $p_\theta(y|x)$.

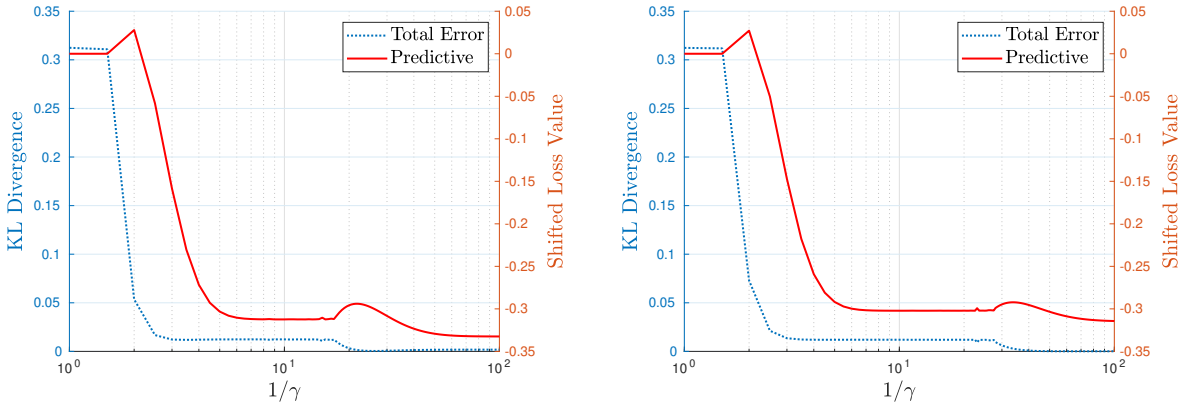
The results are shown in Figure 4.1. We compare the proposed bound (Theorem 4.3.1) to the existing mutual information bounds (denoted as "MI bound") in terms of the generalization error gap. The gap is calculated by the absolute difference of the accuracy rates from testing and training data. We run the testing phase for 200 times and take average of the obtained generalization error gap. The results shown in the figure clearly demonstrate that the proposed new technique can significantly improve the tightness over the mutual information-based bound and this observation holds for all four cases of m. It is worthwhile to highlight that for $1/\gamma < 5$, the proposed bound captures the increase in error gap while the compared bound failed to.

4.5.2 Generalization Error Minimization through the IB Methods

In Chapter 4.4.3, we demonstrate that in the proposed minimax framework, minimizing the surrogate KL divergence upper bound results in optimizing the predictive IB Lagrangian if the Markov chain $X - Z - \hat{Y}$ is satisfied. Following the setup in the last section, we adopt the same synthetic joint measure and the same class of learning model. Our goal here is to solve the proposed minimax problem through the IB methods as surrogate upper bound and compare the optimal solutions of the true mismatch divergence and the surrogate bound. Most importantly, how close they are when mapped on the parameter γ. In Figure 4.2 we first vary the number of samples in sampling and observe the shape of the surrogate bounds in each case. Observe that when the number of samples are limited ($m = 100$ in this case), as $1/\gamma \to 100$, the total error does not converge to zero but a finite value due to the sampling mismatch. And as the number of samples becomes more and more sufficient, the total error for $\gamma \to$ decreases to an infinitesimal value. In Figure 4.2a, the minimum of the total error is



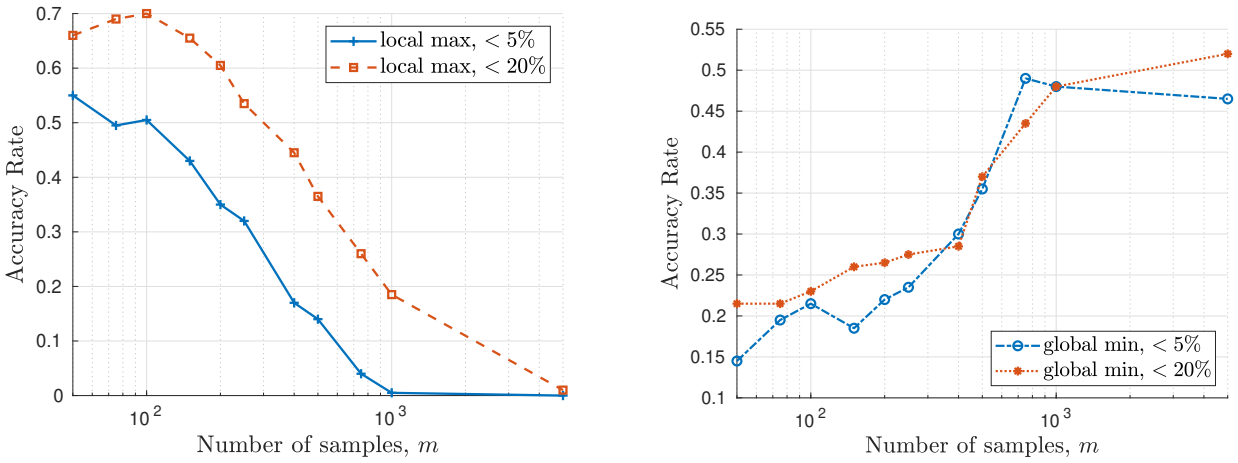
(a) $m = 100$ (b) $m = 500$



(c) $m = 1000$ (d) $m = 5000$

Figure 4.2. Using IB as surrogate loss function in a synthetic standard supervised learning tasks

around $1/\gamma \approx 20$ and about the same location, the red line (Predictive) has a local maxima. The same phenomenon occurs in all m. Based on this observation, we heuristically implement a simple scheme to detect the local maximum whose associated parameters in turns become the candidates of the selected learning model. This heuristic approach corresponds to the common practice in modern machine learn, known as the gradient-based learning [95], where the first-order methods converges to either a local maximum or minimum. We use *findpeak* in MATLAB where we use the function to locate the $\hat{\gamma}_{IB}$ that attains a local maxima to form a candidate set Γ_{IB}. Then we compare each candidate model $\hat{\gamma}_{IB}$ for the predictive IB bound to that achieving the global minima for the total error, denoted as γ^*_{min}. For any $\hat{\gamma}_{IB} \in \Gamma_{IB}$, if the difference $|\hat{\gamma}_{IB} - \gamma^*_{\text{min}}|$ is within a predetermined threshold (in percentage



(a) Local maxima as candidates

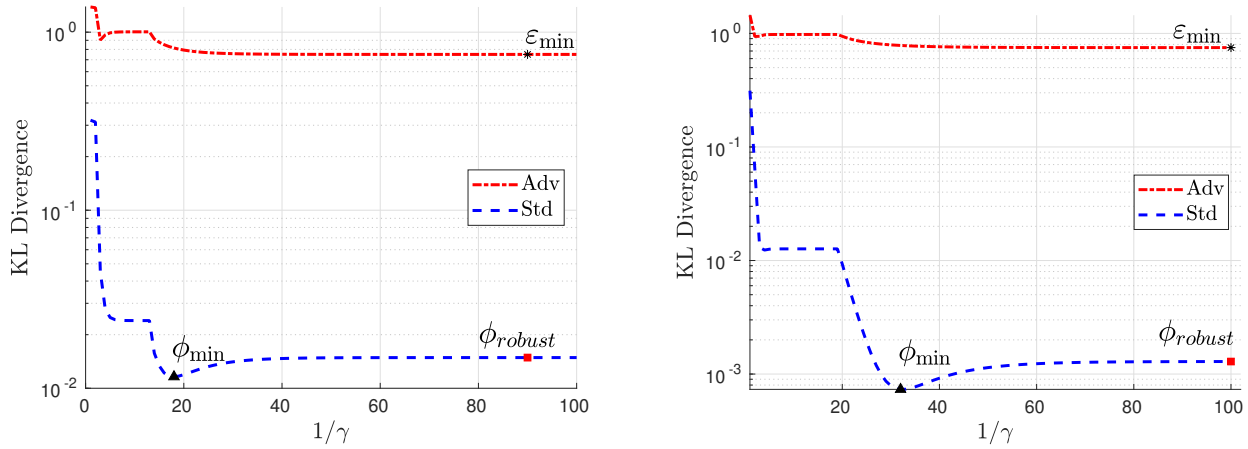
(b) Global minima as candidates

Figure 4.3. Evaluation of the Predictive IB Surrogate Bounds

of a range of γ), then we claim it is accurate. In Figure 4.3, we show that even with this heuristic method, for a threshold that $< 20\%$ of the range of γ, the candidate set that is formed by finding the local maxima reaches about 70% accuracy and around 55% accuracy if further restricting the thresholds to $< 5\%$. However, this approach works well only when the number of samples is limited. We further show another more intuitive approach that simply compares the global minima of the predictive IB bound. As shown in Figure 4.3b, the second approach improves as the number of samples increases. The two results hint a transition of the usefulness that favors the local maxima approach toward the global minima method. Remarkably, if solving the predictive IB Lagrangian with gradient descent, then since this common practice cannot differentiate between local minima and local maxima, it is possible to obtain these specific models. This result sheds light on why learning through the IB methods has been successful at least in supervised learning case. As a final remark, this result can connect to the more general encoder-decoder architecture-based models as the Markov chain $X - Z - \hat{Y}$ holds in this class of architectures.

4.5.3 Adversarial Learning Tasks

The last simulation result for this chapter is the more challenging adversarial learning scenario. For simplicity, we consider a classification task, also we limit the focus on a specific



(a) $m = 100, \alpha = 0.75$ (b) $m = 2000, \alpha = 0.75$

Figure 4.4. Accuracy-Robustness Trade-off in the Proposed Framework

type of adversary who has "read-only" access to the training data and aims at worsening the performance of a learning model by maximizing the divergence mismatch. Note that if given "read-write" access to an adversary, then this corresponds to the scenario of the poisoning attack in literature. Continuing the "read-only" adversary settings, we assume that the learner is aware of the potential presence of the adversary, then intuitively, the learner will avoid fitting the training data completely (overfitting). Here we provide simulation results on some simple but illuminating examples in the following.

First, we explain the adversary's capability in detail. From the proposed minimax framework, which is in fact from a learner's perspective, the learner will pick the worst-case distribution mismatch to mimic the presence of an adversary withing the bounded-divergence controlled through a threshold α, i.e., $D_{KL}[p_\zeta(x,y)||p_\phi(x,y)] \leq \alpha$, where ζ corresponds to the distribution from an adversary while ϕ denotes the joint training distribution. In practice, the learner then needs to find the worst-case attack within the bounded range then find the best hypothesis class θ to achieve optimal robustness against the potential adversary. Note that this procedure is in close resemblance to the best defense strategy today known as the adversarial training [96]. Fortunately, due to the simplicity of the synthetic dataset we focused on for now, we can brute-force search over the joint probability simplex of the worst-case adversarial joint distribution $p_\zeta(x,y)$ by dividing the simplex into grids.

Then, for the learner, in selecting the hypothesis space or the learning models, we again reuse the predictive IB model as elaborated in 4.5.1. The learner also has finite samples of size m, and hence a training distribution $p_\phi(x, y)$. And a predictor is chosen in the following criterion to learn from the m available samples: given a fixed trade-off parameter $\gamma \in [0.01, 1.0]$, we randomly initialization an IB solver for 200 times and pick the set of conditional probability $p_{z|x}$ that attains the highest $I(X; \hat{Y})$ then obtain the associated Bayes' decoder as the predictor.

Our goal is to compare the solution from the proposed minimax framework and provide insights on the challenging adversary learning problems. Therefore, we compare two types of solutions of the proposed minimax problem, corresponding to the standard (Std.) and adversary scenarios (Adv.). The results are shown in Figure 4.4. The threshold is set to $\alpha = 0.75$ and we compare two sets of different training distribution with number of samples $m_1 = 100, m_2 = 2000$. The label $\phi_{\min}$ and $\varepsilon_{\min}$ are the trade-off parameters $\gamma^*_\phi, \gamma^*_\zeta$ that each attains the minimum loss $D_{KL}[\zeta||\theta(\phi_{\min})]$, $D_{KL}[\zeta||\theta(\zeta_{\min})]$ respectively. Then the results provide insights on the recent empirical discovery that after being adversarially trained, a learning model's performance with clean data degrades. This phenomenon, known as the accuracy-robustness trade-off can be explained through our results, even in this simple and limited case. When the learner's goal is to make the model robust to distribution mismatch, then $\varepsilon_{\min}$ is the optimal choice. However, when tested with clean samples, that is, mapping the γ^*_ε to the blue line and hence gives ϕ_{robust}, there is an increase of the divergence loss as compared to the optimal $\phi_{\min}$ in standard learning counterpart. Quantitatively, we can measure the percentage of increase of the extra divergence as $\phi_{\text{robust}}/\phi_{\min} - 1$. In Figure 4.4a the divergence increases 28.15% more than the optimal $\phi_{\min}$ while in Figure 4.4b, the divergence increases 75.85% more than the optimal divergence. This result implies that the challenging adversarial learning problem can be studied through distribution mismatch, as in the proposed framework. As a remark, due to the limitation of computation resources, we are limited to this small-scale problem and restricted hypothesis class, but from our results, it is also clear that more insights are needed before involving more complex, non-linear models or large-scale dataset. This direction of generalization will be left as future works.

5. CONCLUSION AND FUTURE WORKS

5.1 Concluding Remarks

To conclude, in this dissertation, we studied two specific non-convex information-theoretic optimization problems, the IB and PF, that recently attracted significant attention for their successful application in a variety of representation learning or unsupervised clustering problems. Different from previous studies, we approach the two problems from an optimization-mathematics perspective, which is based on the recent breakthrough in non-convex optimization with splitting methods. We propose two types of algorithms in solving the IB and PF problems and prove the convergence and linear rate of convergence of them. In proving the results, we show that the key is the smoothness conditions that allows the application of the powerful KŁ inequality. Based on the theoretic results, new algorithms for both IB and PF are developed. The new solvers simplify the design and relax the convergence conditions for existing solvers and stand out from other greedy algorithm-based solvers as the new ones can handle both random and deterministic mappings. Moreover, we empirically showed that the new solvers can better characterize the relevance-complexity trade-off for IB and privacy-utility trade-off for PF compared to existing solvers.

In response to the recent trend on harvesting the improved performance with learning with multi-modal data. We generalize the proposed splitting methods-based algorithms to MvIB and MsPF, which extends the well-known advantage of splitting methods in convex setting to non-convex optimization context. Inspired by the two extremes in multi-modal representation learning, where on one end there is abundance of representation overlap whereas on the other end there is limited overlap. We developed two distinct information-theoretic formulations and propose two algorithms catering to the two opposite scenarios, in sharp contrast to existing heuristic-based methods. By extending the convergence analysis results for two-block non-convex splitting methods to multi-block consensus generalization, we again proved the linear rate of convergence of the proposed algorithms for both the MvIB and MsPF. We empirically evaluated the new solvers and showed that they avoid the exponential growth of the dimensional complexity in optimal view-merging approach and outperform the state-of-the-art DNN method, which finds representation consensus through based on black-

box neural networks. The results demonstrate the benefits of the new information-theoretic formulations and shed light on the dependency of representation overlap and multi-modal learning gain.

Beyond fitting a given training distribution, by viewing the generalization error in learning problems as distribution mismatch, we propose new framework that formulates the analysis into a minimax problem which can take the standard learning and the more recent adversarial learning tasks into account. In solving the minimax problem, we develop new surrogate upper bound based on the Pythagorean theorem which gives tighter upper bounds than existing ones. Moreover, we demonstrate that the proposed Pythagorean theorem-based technique applies to a variety of existing bound-tightening techniques and extends to a broader class of distribution mismatch metrics, i.e., the α-information through the Generalized Pythagorean theorem. Our results further connect to the IB methods for learning that extend to the celebrated encoder-decoder architecture in modern machine learning practices. Using SDPI, we show that solving the predictive IB Lagrangian is equivalent to optimizing over a distribution mismatch caused by sampling process or a potential adversary. Empirically, we compare the derived upper bounds to existing methods in a standard supervised learning task with synthetic data, our results give significantly tighter bound compared to the recent input-output mutual information bounds. In addition, when adopting predictive IB as the surrogate upper bound in solving the proposed minimax problem, we find that it can produce robust learning models that are close to the optimal choice. Lastly, again based on the perspective of distribution mismatch, we evaluate the proposed framework on an adversarial learning task and provide insights on the accuracy-robustness trade-off discovered recently.

5.2 Future works

Beyond the results obtained in this dissertation, we identify the open challenges and directions for future exploration in the following.

5.2.1 Application to Graph-Based Geometric Clustering Problems

In literature, IB methods have already been applied to clustering and graph-based multivariate problems [6], [97], but recently due to the surging interests in the graph neural network (GNN) in modern machine learning research, driven by the state-of-the-art performance in various learning problems, the IB methods for graph-based applications regain attention lately. Equipped with modern optimization and machine learning advances, the graph-based IB methods today are adopted to extract latent features from highly structured or geometric-dependent data. Applications following this resurgence include natural language processing, geometric clustering and adversarial machine learning [98]–[100]. Meanwhile, ADMM has recently been extended to solve index programming problem [101], which applies to clustering. On the other hand, in clustering with (Mv)IB methods, or (Ms)PF alike, restrictions on deterministic mappings are often needed which makes the optimization problem non-smooth. In this sense, the extension for ADMM to index programming can potentially generalize our theoretic results to clustering problem as the KŁ inequality also applies to non-smooth problems [36], [66].

5.2.2 Generalization to Continuous Settings

While we limit our focus on discrete random variables in this dissertation mostly but a natural question that follows is how much results derived from the discrete settings still hold when applied to continuous settings? In this direction, the vector form of (conditional) probabilities needs to be "parameterized", a commonly adopted approach in density or entropy estimation research [102]–[105], and since the sub-objective functions are in fact a combination of entropy and conditional entropy functions, or in continuous cases the differential entropy. If some smoothness conditions are imposed, so that the differential entropy are well-defined, then the primal and augmented variable updates in discrete settings can be generalized to parameter updates by treating the decoupled sub-objectives as separate entropy estimation problems. However, the update algorithms would inevitably need to be modified to stochastic gradient descent (SGD [106]). Fortunately, SGD has been intensively studied in machine learning research, so there are off-the-shelf optimizers available for the

convenience of empirical evaluation and further applies to splitting methods recently [107]–[110]. However, as the theoretic study of SGD is still an active field, not only the theoretic analysis but the practical implementation are significantly more challenging.

5.2.3 Adversarial Generalization Error Analysis

In Chapter 4 we present a new information-theoretic framework for the generalization error analysis from the perspective of distribution mismatch. We develop new techniques to solve the proposed minimax problem. While for standard learning tasks the new technique gives tighter bounds then existing ones, when applied the same technique to more challenging adversarial settings, we find that the proposed upper bound only holds for a restricted region in terms of the level of fitting for an adopted learning model. Based on our insights from distribution mismatch, we plan to study slightly more complicated, but well-studied learning models, such the support vector machine [111] and then generalize to DNN and more recent architectures. In addition, while for convenience we limit the focus on adversaries with "read-only" access to training data, we also expect to apply the results we have from the relatively passive setting to "read-write" access adversaries in an online learning scenario which corresponds to the poisoning attack [112], [113] that attracts significant attention in modern machine learning research recently.

REFERENCES

[1] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, no. 12, p. 1181, 2019, ISSN: 1099-4300. DOI: 10.3390/e21121181.

[2] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *CoRR*, vol. abs/1612.00410, 2016. arXiv: 1612.00410.

[3] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *2014 IEEE Information Theory Workshop (ITW 2014)*, IEEE, 2014, pp. 501–505.

[4] A. Bardera, J. Rigau, I. Boada, M. Feixas, and M. Sbert, "Image segmentation using information bottleneck method," *IEEE Transactions on Image Processing*, vol. 18, no. 7, pp. 1601–1612, 2009. DOI: 10.1109/TIP.2009.2017823.

[5] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[6] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, and K. Müller, Eds., MIT Press, 2000, pp. 617–623.

[7] S. Hu, Z. Shi, and Y. Ye, "Dmib: Dual-correlated multivariate information bottleneck for multiview clustering," *IEEE Transactions on Cybernetics*, pp. 1–15, 2020. DOI: 10.1109/TCYB.2020.3025636.

[8] C. Xu, D. Tao, and C. Xu, "Large-margin multi-viewinformation bottleneck," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1559–1572, 2014. DOI: 10.1109/TPAMI.2013.2296528.

[9] Q. Wang, C. Boudreau, Q. Luo, P.-N. Tan, and J. Zhou, "Deep multi-view information bottleneck," in *Proceedings of the 2019 SIAM International Conference on Data Mining*, SIAM, 2019, pp. 37–45.

[10] T.-Y. Liu and I.-H. Wang, *Robust privatization with non-specific tasks and the optimal privacy-utility tradeoff*, 2021. arXiv: 2010.10081 [cs.IT].

[11] M. Federici, A. Dutta, P. Forré, N. Kushman, and Z. Akata, "Learning robust representations via multi-view information bottleneck," in *International Conference on Learning Representations*, 2020.

[12] H. Hsu, N. L. Martinezgil, M. Bertran, G. Sapiro, and F. Calmon, "A survey on privacy from statistical, information and estimation-theoretic views," *IEEE BITS the Information Theory Magazine*, pp. 1–1, 2021. DOI: 10.1109/MBITS.2021.3108124.

[13] M. Lopuhaä-Zwakenberg, *The privacy funnel from the viewpoint of local differential privacy*, 2020. arXiv: 2002.01501 [cs.CR].

[14] S. Shalev-Shwartz, *Understanding machine learning : from theory to algorithms*, eng. Cambridge: Cambridge University Press, 2014, ISBN: 1-107-29801-6.

[15] O. Shamir, S. Sabato, and N. Tishby, "Learning and generalization with the information bottleneck," *Theoretical Computer Science*, vol. 411, no. 29-30, pp. 2696–2711, 2010.

[16] M. Vera, P. Piantanida, and L. R. Vega, "The role of the information bottleneck in representation learning," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 1580–1584. DOI: 10.1109/ISIT.2018.8437679.

[17] D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, 2020. DOI: 10.1109/TIT.2019.2945779.

[18] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[19] T. Wu, I. Fischer, I. L. Chuang, and M. Tegmark, "Learnability for the information bottleneck," *CoRR*, vol. abs/1907.07331, 2019. arXiv: 1907.07331.

[20] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, "Information bottleneck for gaussian variables," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. K. Saul, and B. Schölkopf, Eds., MIT Press, 2004, pp. 1213–1220.

[21] J. Nocedal, *Numerical optimization*, eng, 2nd ed., ser. Springer series in operations research. New York: Springer, 2006, ISBN: 0387303030.

[22] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972. DOI: 10.1109/TIT.1972.1054855.

[23] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *information and computation*, vol. 132, no. 1, pp. 1–63, 1997.

[24] K. Nakagawa, Y. Takei, S.-i. Hara, and K. Watabe, "Analysis of the convergence speed of the arimoto-blahut algorithm by the second-order recurrence formula," *IEEE Transactions on Information Theory*, pp. 1–1, 2021. DOI: 10.1109/TIT.2021.3095406.

[25] F. Bayat and S. Wei, "Information bottleneck problem revisited," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2019, pp. 40–47. DOI: 10.1109/ALLERTON.2019.8919752.

[26] D. Strouse and D. J. Schwab, "The deterministic information bottleneck," *Neural computation*, vol. 29, no. 6, pp. 1611–1630, 2017.

[27] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, Apr. 2017. DOI: 10.1080/01621459.2017.1285773.

[28] M. Igl, K. Ciosek, Y. Li, S. Tschiatschek, C. Zhang, S. Devlin, and K. Hofmann, "Generalization in reinforcement learning with selective noise injection and information bottleneck," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[29] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*, PMLR, 2020, pp. 7836–7846.

[30] N. Ding and P. Sadeghi, "A submodularity-based clustering algorithm for the information bottleneck and privacy funnel," in *2019 IEEE Information Theory Workshop (ITW)*, 2019, pp. 1–5. DOI: 10.1109/ITW44776.2019.8989355.

[31] M. Narasimhan and J. A. Bilmes, "A submodular-supermodular procedure with applications to discriminative structure learning," *arXiv preprint arXiv:1207.1404*, 2012.

[32] I. Fischer, "The conditional entropy bottleneck," *Entropy*, vol. 22, no. 9, 2020, ISSN: 1099-4300. DOI: 10.3390/e22090999.

[33] S. Boyd, N. Parikh, and E. Chu, *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[34] T.-H. Huang and A. el Gamal, "A provably convergent information bottleneck solution via admm," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 43–48. DOI: 10.1109/ISIT45174.2021.9518141.

[35] T. Zhang and Z. Shen, "A fundamental proof of convergence of alternating direction method of multipliers for weakly convex optimization," *Journal of Inequalities and Applications*, vol. 2019, no. 1, 2019, ISSN: 1029-242X. DOI: 10.1186/s13660-019-2080-0.

[36] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Mathematical Programming*, vol. 116, no. 1, pp. 5–16, 2009.

[37] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-Łojasiewicz inequality," *Mathematics of operations research*, vol. 35, no. 2, pp. 438–457, 2010.

[38] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013, ISSN: 0025-5610. DOI: 10.1007/s10107-011-0484-9.

[39] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016, ISSN: 1052-6234. DOI: 10.1137/140990309.

[40] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd, "First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems," *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2131–2151, 2018.

[41] M. Chao, D. Han, and X. Cai, "Convergence of the peaceman-rachford splitting method for a class of nonconvex programs," *Numerical Mathematics: Theory, Methods and Applications*, vol. 14, no. 2, pp. 438–460, 2021, ISSN: 2079-7338.

[42] B. He and X. Yuan, "On non-ergodic convergence rate of douglas–rachford alternating direction method of multipliers," *Numerische Mathematik*, vol. 130, no. 3, pp. 567–577, 2015, ISSN: 0029-599X. DOI: 10.1007/s00211-014-0673-6.

[43] A. Themelis and P. Patrinos, "Douglas–rachford splitting and admm for nonconvex optimization: Tight convergence results," *SIAM Journal on Optimization*, vol. 30, no. 1, pp. 149–181, 2020, ISSN: 1052-6234. DOI: 10.1137/18m1163993.

[44] T.-H. Huang, A. E. Gamal, and H. E. Gamal, "On the multi-view information bottleneck representation," *arXiv preprint arXiv:2202.02684*, 2022.

[45] Y. Wang, W. Yin, and J. Zeng, "Global convergence of admm in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, 2019, ISSN: 0885-7474. DOI: 10.1007/s10915-018-0757-z.

[46] K. Guo, D. Han, and X. Yuan, "Convergence analysis of douglas–rachford splitting method for "strongly + weakly" convex programming," *SIAM Journal on Numerical Analysis*, vol. 55, no. 4, pp. 1549–1577, 2017, ISSN: 0036-1429. DOI: 10.1137/16m1078604.

[47] Z. Jia, X. Gao, X. Cai, and D. Han, "Local linear convergence of the alternating direction method of multipliers for nonconvex separable optimization problems," *Journal of Optimization Theory and Applications*, vol. 188, no. 1, pp. 1–25, 2021, ISSN: 0022-3239. DOI: 10.1007/s10957-020-01782-y.

[48] K. Kurdyka, "On gradients of functions definable in o-minimal structures," eng, *Annales de l'institut Fourier*, vol. 48, no. 3, pp. 769–783, 1998.

[49] G. Li and T. K. Pong, "Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods," *Foundations of Computational Mathematics*, vol. 18, no. 5, pp. 1199–1232, 2018, ISSN: 1615-3375. DOI: 10.1007/s10208-017-9366-8.

[50] S. Sun, "A survey of multi-view machine learning," *Neural computing and applications*, vol. 23, no. 7, pp. 2031–2038, 2013.

[51] Y. Yang and H. Wang, "Multi-view clustering: A survey," *Big Data Mining and Analytics*, vol. 1, no. 2, pp. 83–107, 2018. DOI: 10.26599/BDMA.2018.9020003.

[52] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*, PMLR, 2015, pp. 1083–1092.

[53] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1863–1883, 2019. DOI: 10.1109/TKDE.2018.2872063.

[54] K. Zhan, F. Nie, J. Wang, and Y. Yang, "Multiview consensus graph clustering," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1261–1270, 2019. DOI: 10.1109/TIP.2018.2877335.

[55] Y. Gao, S. Gu, L. Xia, and Y. Fei, "Web document clustering with multi-view information bottleneck," in *2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06)*, 2006, pp. 148–148. DOI: 10.1109/CIMCA.2006.232.

[56] V. Vapnik, *The Nature of Statistical Learning Theory*, eng, 2nd ed. 2000., ser. Information Science and Statistics. New York, NY: Springer New York, 2000.

[57] O. Bousquet and A. Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, pp. 499–526, 2002.

[58] M. S. Masiha, A. Gohari, M. H. Yassaee, and M. R. Aref, "Learning under distribution mismatch and model misspecification," in *2021 IEEE International Symposium on Information Theory (ISIT)*, 2021, pp. 2912–2917. DOI: 10.1109/ISIT45174.2021.9517732.

[59] A. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[60] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, 2020. DOI: 10.1109/JSAIT.2020.2991139.

[61] H. Harutyunyan, M. Raginsky, G. Ver Steeg, and A. Galstyan, "Information-theoretic generalization bounds for black-box learning algorithms," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[62] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *Lecture Notes for ECE563 (UIUC) and*, vol. 6, no. 2012-2016, p. 7, 2014.

[63] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 824–839, 2020. DOI: 10.1109/JSAIT.2020.3040992.

[64] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International conference on machine learning*, PMLR, 2018, pp. 531–540.

[65] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience, 2006, ISBN: 0471241954.

[66] J. Bolte, S. Sabach, and M. Teboulle, "Nonconvex lagrangian-based optimization: Monitoring schemes and global convergence," *Mathematics of Operations Research*, vol. 43, no. 4, pp. 1210–1232, 2018.

[67] K. Sricharan, R. Raich, and A. O. Hero, "Estimation of nonlinear functionals of densities with confidence," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4135–4159, 2012. DOI: 10.1109/TIT.2012.2195549.

[68] Y. Han, J. Jiao, T. Weissman, and Y. Wu, "Optimal rates of entropy estimation over lipschitz balls," *The Annals of Statistics*, vol. 48, no. 6, pp. 3228–3250, 2020.

[69] I. Sason, "On reverse pinsker inequalities," *CoRR*, vol. abs/1503.07118, 2015. arXiv: 1503.07118.

[70] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 2013, pp. 429–438. DOI: 10.1109/FOCS.2013.53.

[71] H. H. Bauschke, J. Bolte, and M. Teboulle, "A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications," *Mathematics of Operations Research*, vol. 42, no. 2, pp. 330–348, 2017.

[72] Y. Nesterov, *Lectures on Convex Optimization*, eng, 2nd ed. 2018., ser. Springer Optimization and Its Applications, 137. Cham: Springer International Publishing, 2018, ISBN: 3-319-91578-9.

[73] A. E. Parker, A. G. Dimitrov, and T. Gedeon, "Symmetry breaking in soft clustering decoding of neural codes," *IEEE Transactions on Information Theory*, vol. 56, no. 2, pp. 901–927, 2010.

[74] T. Gedeon, A. E. Parker, and A. G. Dimitrov, "The mathematical structure of information bottleneck methods," *Entropy*, vol. 14, no. 3, pp. 456–479, 2012. DOI: 10.3390/e14030456.

[75] G. Li and T. K. Pong, "Global convergence of splitting methods for nonconvex composite optimization," *SIAM Journal on Optimization*, vol. 25, no. 4, pp. 2434–2460, 2015, ISSN: 1052-6234. DOI: 10.1137/140998135.

[76] T. Wu and I. Fischer, "Phase transitions for the information bottleneck in representation learning," in *International Conference on Learning Representations*, 2020.

[77] D. P. Bertsekas, *Nonlinear programming*, eng, 2nd ed. Belmont, Mass.: Athena Scientific, 1999, ISBN: 9781886529007.

[78] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, 2020, ISSN: 1472-6947. DOI: 10.1186/s12911-020-1023-5.

[79] D. Dua and C. Graff, *UCI machine learning repository*, 2017. [Online]. Available: http://archive.ics.uci.edu/ml.

[80] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, ser. COLT' 98, Madison, Wisconsin, USA: Association for Computing Machinery, 1998, pp. 92–100, ISBN: 1581130570. DOI: 10.1145/279943.279962.

[81] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[82] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015.

[83] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2.

[84] A. Kurakin, I. Goodfellow, S. Bengio, *et al.*, *Adversarial examples in the physical world*, 2016.

[85] F. d. P. Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities for input constrained additive noise channels," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1879–1892, 2018. DOI: 10.1109/TIT.2017.2782359.

[86] Y. Polyanskiy and Y. Wu, "Dissipation of information in channels with input constraints," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 35–55, 2016. DOI: 10.1109/TIT.2015.2482978.

[87] Z. Wang, S.-L. Huang, E. E. Kuruoglu, J. Sun, X. Chen, and Y. Zheng, "PAC-bayes information bottleneck," in *International Conference on Learning Representations*, 2022.

[88] J. Shawe-Taylor, *Kernel methods for pattern analysis*, eng. Cambridge: Cambridge University Press, 2004, ISBN: 1-107-14456-6.

[89] M. Raginsky, "Strong data processing inequalities and Φ-sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, 2016. DOI: 10.1109/TIT.2016.2549542.

[90] Z. Piran, R. Shwartz-Ziv, and N. Tishby, "The dual information bottleneck," *arXiv preprint arXiv:2006.04641*, 2020.

[91] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[92] S. Boucheron, *Concentration inequalities : a nonasymptotic theory of independence*, eng, 1st ed. Oxford: Oxford University Press, 2013, ISBN: 9780199535255.

[93] H. L. Royden, *Real analysis*, eng, 4th ed. Boston: Prentice Hall, 2010.

[94] T. van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014. DOI: 10.1109/TIT.2014.2320500.

[95] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[96] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[97] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby, "Multivariate information bottleneck," *arXiv preprint arXiv:1301.2270*, 2013.

[98] D. Beck, G. Haffari, and T. Cohn, "Graph-to-sequence learning using gated graph neural networks," *arXiv preprint arXiv:1806.09835*, 2018.

[99] D. Strouse and D. J. Schwab, "The information bottleneck and geometric clustering," *Neural Computation*, vol. 31, no. 3, pp. 596–612, 2019, ISSN: 0899-7667. DOI: 10.1162/neco_a_01136.

[100] T. Wu, H. Ren, P. Li, and J. Leskovec, "Graph information bottleneck," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 437–20 448, 2020.

[101] B. Wu and B. Ghanem, "ℓ_p p-box admm: A versatile framework for integer programming," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1695–1708, 2019. DOI: 10.1109/TPAMI.2018.2845842.

[102] G. Ciuperca, V. Girardin, and L. Lhote, "Computation and estimation of generalized entropy rates for denumerable markov chains," *IEEE Transactions on Information Theory*, vol. 57, no. 7, pp. 4026–4034, 2011. DOI: 10.1109/TIT.2011.2133710.

[103] Z. Goldfeld, K. Greenewald, J. Niles-Weed, and Y. Polyanskiy, "Convergence of smoothed empirical measures with applications to entropy estimation," *IEEE Transactions on Information Theory*, vol. 66, no. 7, pp. 4368–4391, 2020. DOI: 10.1109/TIT.2020.2975480.

[104] A. Mokkadem, "Estimation of the entropy and information of absolutely continuous random variables," *IEEE Transactions on Information Theory*, vol. 35, no. 1, pp. 193–196, 1989. DOI: 10.1109/18.42194.

[105] K. Sricharan, D. Wei, and A. O. Hero, "Ensemble estimators for multivariate entropy estimation," *IEEE Transactions on Information Theory*, vol. 59, no. 7, pp. 4374–4388, 2013. DOI: 10.1109/TIT.2013.2251456.

[106] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, Springer, 2010, pp. 177–186.

[107] W. Zhong and J. Kwok, "Fast stochastic alternating direction method of multipliers," in *International Conference on Machine Learning*, PMLR, 2014, pp. 46–54.

[108] B. Sirb and X. Ye, "Decentralized consensus algorithm with delayed and stochastic gradients," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1232–1254, 2018.

[109] X. Li and F. Orabona, "On the convergence of stochastic gradient descent with adaptive stepsizes," in *The 22nd International Conference on Artificial Intelligence and Statistics*, PMLR, 2019, pp. 983–992.

[110] A. Barakat and P. Bianchi, "Convergence and dynamical behavior of the adam algorithm for nonconvex stochastic optimization," *SIAM Journal on Optimization*, vol. 31, no. 1, pp. 244–274, 2021.

[111] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[112] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," *Advances in neural information processing systems*, vol. 30, 2017.

[113] Y. Ma, X. Zhu, and J. Hsu, "Data poisoning against differentially-private learners: Attacks and defenses," *arXiv preprint arXiv:1903.09860*, 2019.

VITA

Teng-Hui Huang earned a bachelor's degree in Electrical and Engineering in 2015 from National Taiwan University, Taipei, Taiwan, and a master's degree in Engineering in 2017 from the Graduate Institute of Communication Engineering of National Taiwan University, Taipei, Taiwan.