A Linearly Convergent Douglas-Rachford Splitting Solver for Markovian Information-Theoretic Optimization Problems

Teng-Hui Huang, Student Member, IEEE, Aly El Gamal, Senior Member, IEEE, and Hesham El Gamal, Fellow, IEEE

Abstract—In this work, we propose solving the Information Bottleneck (IB) and Privacy Funnel (PF) problems with Douglas-Rachford Splitting methods (DRS). We study a general Markovian information-theoretic Lagrangian that includes IB and PF into a unified framework. We prove the linear convergence of the proposed solvers using the Kurdyka-Łojasiewicz inequality. Moreover, our analysis is beyond IB and PF and applies to any convex-weakly convex pair objectives. Based on the results, we develop two types of linearly convergent IB solvers, with one improves the performance of convergence over existing solvers while the other can be independent to the relevance-compression trade-off. Moreover, our results apply to PF, yielding a new class of linearly convergent PF solvers. Empirically, the proposed IB solvers IB obtain solutions that are comparable to the Blahut-Arimoto-based benchmark and is convergent for a wider range of the penalty coefficients than existing solvers. For PF, our non-greedy solvers can characterize the privacy-utility trade-off better than the clustering-based greedy solvers.

Index Terms—Mutual information, information bottleneck (IB), privacy funnel (PF), source coding, convergence, iterative algorithm, optimization methods, gradient methods, Lagrangian functions.

I. INTRODUCTION

RECENTLY, adopting information-theoretic metrics as optimization objectives drew significant attention from machine learning and data science communities. Among which, the information bottleneck (IB) methods [1], studying the complexity-relevance trade-off of representations, have been applied to a wide range of applications including representation learning, classification and clustering problems with impressive performance [2], [3], [4]. However, the advances of machine learning also bring new challenges. Collecting a large amount of data burdens the conventional centralized learning approach; On the other hand, the leakage of sensitive information becomes a major concern for both the end users and industries [5], [6], [7]. Closely related to this, the privacy funnel (PF) [8], another information-theoretic optimization problem, has been adopted in optimization objectives as it characterizes the trade-off between the leakage of sensitive information and the utility of observations.

A. Information Bottleneck and Privacy Funnel

We start with a brief review of the IB and PF problems. Given the joint probability of the observations $X$ and targets $Y$, the IB methods aim to find a representation $Z$ of the observations that is minimal in expression complexity but retains most relevance to the target. In IB, this goal is formulated as a constrained optimization problem [1]:

$$\minimize_{p(z|x) \in \Omega} I(Z; X),$$
subject to: $I(Z; Y) \geq I_0,$
$$\sum_z p(z|x) = 1, \forall x \in X,$$
$$Y \rightarrow X \rightarrow Z \ \text{Markov chain}, \ (1)$$

where the constant $I_0 \geq 0$ controls the trade-off between the mutual information $I(Z; X)$ and $I(Z; Y)$; $\Omega$ is a compound probability simplex, and the variable to optimize is the conditional probability $p(Z|X)$. In solving the IB problem, (1) can be relaxed as the following IB Lagrangian:

$$L_{IB} := \gamma I(Z; X) - I(Z; Y), \ (2)$$

where the multiplier $\gamma$ controls the trade-off between the two competing objectives.

The PF problem considers the opposite scenario: denote $Y$ as the sensitive information, $X$ the public information and $Z$ the observation, the goal of PF is to design an information assignment $p(Z|X)$ such that the utility of $Z$ to the public information is maximized but minimal sensitive information is revealed from $Z$. This goal satisfies the Markov chain $Y \rightarrow X \rightarrow Z$ where the utility and privacy leakage are measured by the mutual information $I(Z; X)$ and $I(Z; Y)$ respectively. Then to find the optimal assignment, one minimizes $I(Z; Y)$ but maximizes $I(Z; X)$ over $p(Z|X)$ with a known $p(X, Y)$.
Through the Lagrange multipliers, this optimization problem can be written as the following PF Lagrangian:

$$\mathcal{L}_{PF} := \beta I(Z; Y) - I(Z; X), \quad (3)$$

where $\beta$ denotes the multiplier. Both IB and PF problems are known to be non-convex therefore are difficult to solve [1], [8]. A practical approach to minimize the IB and PF Lagrangian is to fix a trade-off parameter (the multiplier) and solve the relaxed problem. Then by varying a range of the trade-off parameter, one can obtain a collection of $p(Z|X)$, hence pairs of $I(Z; X), I(Z; Y)$.

To evaluate solutions obtained from both the IB and PF problems, one plots each pair of $I(Z; X), I(Z; Y)$ on the $x, y$-axes respectively. The plot is called the information plane [1], [9]. On the information plane, the best $I(Z; Y)$ given a $I(Z; X)$ is revealed. For a fixed $I(Z; X)$, a higher $I(Z; Y)$ is better in the IB whereas a lower $I(Z; Y)$ is preferred in PF. Then the frontier formed from the set of best $I(Z; Y)$ over a range of $I(Z; X)$ characterizes the relevance-complexity trade-off for IB, and the privacy-utility trade-off for PF [1], [6], [8], [9]. This frontier is known as the Pareto-frontier [3], [10] and can be used to evaluate the performance of IB and PF solvers.

### B. IB and PF Solvers

The formulations of IB and PF are quite general. Hence a variety of solvers have been developed [1], [2], [5], [9], [11], [12], [13], [14], [15], [16]. Inspired by the Blahut-Arimoto (BA) algorithm in rate distortion theory, a set of self-consistent techniques. For instance, the deep variational IB (DVIB) [2] and the exact posterior [19], [20]. Note that for discrete IB scenarios, the variational inference-based method reduces to the BA-based algorithm (4c).

While the above problem is convex with respect to $p(z|x)$, the feasible set is more restricted so only a limited number of points on the information plane are found.

Following the recent innovations in non-convex optimization with splitting methods, [15] empirically shows that the IB problem can be solved with the alternating direction method of multipliers (ADMM) [18] by separating the IB Lagrangian into three terms of (conditional) Shannon entropy functions $H(Z), H(Z|X)$ and $H(Z|Y)$. In our earlier work [16], we simplify the method into two terms, significantly reduce the complexity and further prove the convergence.

The aforementioned solvers are susceptible to scalability when the cardinalities of $X$ and $Y$ are large. This issue has been recently addressed through variational inference techniques. For instance, the deep variational IB (DVIB) [2] and a deep variational solver for the PF [5] have been developed. However, the variational inference methods are approximate in nature and will provide exact solutions for the IB or PF Lagrangian only if the variational class of distributions includes the exact posterior [19], [20].

### C. Contributions

In this work, we consider a general class of discrete Markovian information-theoretic optimization problems. This class includes both the IB and PF problems as special cases. We develop a framework for constructing solvers for this class of problems via the Douglas-Rachford splitting (DRS) methods [23]. We construct a unified convergence proof of the proposed set of solvers which guarantees convergence without relying on additional regularization terms; unlike earlier works for ADMM-based solvers [15], [16]. Further, we prove that the rate of convergence for the proposed solvers is locally linear.

The main tool in our convergence analysis is the Kurdyka-Łojasiewicz (KL) inequality [24], [25], [26] which is included in appendices for completeness.

For the IB special case, our approach depends on decomposing the objective function into two sub-objectives via two different alternative approaches. Each approach yields a different convergent solver as described in Section IV. The first solver guarantees convergence by requiring a smaller penalty coefficient, as compared with existing ADMM-based IB solvers, and exploiting the strong-convexity of the sub-objective. The second solver ensures convergence by exploiting Lipschitz smoothness and weak-convexity of one of the sub-objectives, independent of the trade-off parameter $\gamma$ (except for the case where $X$ is obtained from $Y$ through a deterministic mapping).

For the PF special case, we develop a new solver which recovers all the points on the information plane which are found by existing clustering-based greedy algorithms and adds to them more points not reachable by existing solvers. Contrary to the work reported in [5], the decoder part of our solver satisfies the Markovian condition (i.e., it is independent of the sensitive information $Y$). Hence, it can explore the privacy-utility trade-off better in the discrete and Markovian settings.
Empirically, we evaluate the new IB and PF solvers on both synthetic and real-world data. For the IB case, the proposed solvers are shown to be more comprehensive in exploring the information plane as compared with the benchmark BA-based solver. The new IB solvers are also convergent for a broader range of design parameters compared to existing ADMM-based solvers. Similarly, for the PF special case, our numerical results illustrate that the new solver is able to obtain more points on the information plane than existing greedy based solvers [3], [8]. Finally, we show that our solvers are efficient in computing the variational bounds on both the IB and PF problems. In this sense, solving the augmented Lagrangian (9) is a minimizer to the augmented Lagrangian (9) to the original Lagrangian. In this sense, solving the augmented Lagrangian is easier because the equality constraints can be relaxed during optimization [31].

Instead of jointly minimizing (9) over \( p,q,\nu \), splitting methods solve (9) in an alternating fashion. More specifically, splitting methods minimize (9) by repeatedly updating \( p,q \) then \( \nu \). When both sub-objectives \( F,G \) are convex, the convergence of splitting methods are well studied [18], [32]. Furthermore, the rate of convergence of these solvers have been characterized in [33] and [34]. In contrast, there are less results for the cases where \( F(p) + G(q) \) is non-convex. Recently, several works have empirically found that if one of the two sub-objectives is convex and the other weakly convex, then splitting methods can solve this problem effectively [35]. Following the discovery, convergence of splitting methods for objective with this structure has been developed in [21], [22], [27], [36], and [37]. In addition to convergence, the rate of convergence can also be shown to be locally linear through the Kurdyka-Lojasiewicz inequality (see Appendix A-B for details).

Among the general class of splitting methods, a well-known solver is the alternating direction method of multiplier (ADMM) [18]. In [21], ADMM is adopted to solve non-convex quadratic programming with simplex constraints and it is shown to outperform the state-of-the-art. In [38], the ADMM algorithm is used to solve non-convex LASSO faster than known methods with better performance. In [35] the Douglas-Rachford Splitting (DRS) [23] method, where ADMM is a special case of it, is used to find sparse solutions of a linear system [35].

### B. The Proposed Solvers

We propose two solvers for the augmented Lagrangian (9). The first solver, which will be referred to as Solver I in the sequel, is described by the set of update equations:

\[
\nu_{1/2}^{k+1} := \nu^{k} - (1 - \alpha) c \left( A p^{k} - B q^{k} \right),
\]

(10a)

\[
p^{k+1} := \arg \min_{p \in \Omega_p} \mathcal{L}_c(p, q^{k}, \nu^{k+1/2}),
\]

(10b)

\[
\nu^{k+1} := \nu_{1/2}^{k+1} + c \left( A p^{k+1} - B q^{k+1} \right),
\]

(10c)

\[
q^{k+1} := \arg \min_{q \in \Omega_q} \mathcal{L}_c(p^{k+1}, q, \nu^{k+1}),
\]

(10d)

where the superscript \( k \) denotes the iteration counter; \( \nu_{1/2}^{k+1} \) is an intermediate dual variable accounting the fixed-point relaxation [39] and \( \Omega_p, \Omega_q \) are probability simplexes for \( p,q \) respectively; \( \alpha > 0 \) is a relaxation parameter where \( 1 < \alpha < 2 \) is “over-relaxation” while \( 0 < \alpha < 1 \) is “under-relaxation”. We refer to [39] for details. Note that \( \alpha = 1 \) recovers the ADMM whereas \( \alpha = 2 \) corresponds to the Peaceman-Rachford splitting (PRS) [40]. Alternatively, by changing the updating order of (10), we obtain our second DRS solver, denoted as Solver II.

\[
p^{k+1} := \arg \min_{p \in \Omega_p} \mathcal{L}_c(p, q^{k}, \nu^{k})
\]

(11a)
TABLE I

| Reference | Algorithm | Convergence Conditions | Rate of Conv. | Properties of Functions | Linear Constraints |
|-----------|-----------|------------------------|--------------|-------------------------|-------------------|
| Solver I (10) | 0 < α ≤ 2 | c > max{ω_G, \frac{L_p + F}{\nu_A^2}} | locally linear | F: σ_F-strongly convex | A-positive definite |
| Solver II (11) | 0 < α < 2 | c > \frac{α_2 + 1 - \nu}{\nu} * | locally linear | F: convex | A-full row rank |
| Solver II (11) | 0 < α < 2 | c > M_q(\frac{α_2 + 1 - \nu}{\nu}) | locally linear | G: \frac{\nu_2}{\nu} -weakly convex \ G is \ \frac{\nu_2}{\nu}-smooth | A-positive definite |
| Jia et al. [21] | Prox. ADMM (α = 1) | c > \frac{F_α + \sqrt{F_α^2 + 8\sigma_2^2}}{2\nu_2} | locally linear | F: convex | A-positive definite |
| Themelis et al. [22] | 0 < α < 2; 2 ≤ α < 4 | c > L_p; \frac{α}{\nu} < c < \frac{α}{\nu} + 1 | best case sublinear | \ O(1/\sqrt{\nu}) | F: \frac{\nu_2}{\nu}-hypo-cvx; (0 < α < 2) |
|                     |                     |                     |                     | F: \frac{\nu_2}{\nu}-str. cvx. (2 ≤ α < 4) | A = B = 1 |

Remarkably, while Solver I and Solver II seem different only in the order they optimize p, q, when implemented with gradient descent, their first-order necessary conditions are quite different. In particular, the dual ascend connects to the gradient difference of the sub-objective function F for Solver I whereas the connection between the dual variables is on the gradient of G for Solver II.

Moreover, from the perspective of parameter selection, as will be shown in the next section, the rate of convergence of the two solvers depends on the relaxation parameter. Specifically, when 1 ≤ α ≤ 2, the rates are Q-linear while for the case where 0 < α < 1, we can only show R-linear rates, which are weaker than its Q-linear counterpart [41]. However, the strong rate comes with additional assumptions on the linear constraints which limits the problems it applies to. Also, it turns out that the convergence of Solver I and II depends on the choice of the penalty coefficient c [21], [22]. Therefore, knowing the smallest penalty coefficient that assures convergence is essential to the success of applying DRS solvers to non-convex optimization problems.

III. MAIN RESULTS

Our main results are the convergence and rate analysis of the two solvers under three different sets of assumptions (Table I), which serves as a general guideline for applying DRS solvers to non-convex optimization problems beyond IB and PF. For each proposed solver-assumption pair listed in Table I, the rate of convergence is locally linear, using the KL inequality. In particular, we show that the augmented Lagrangian (9) solved with the two algorithms satisfies the KL property with exponent θ = 1/2, hence corresponds to linear rate of convergence [24], [26], [42].

Theorem 1: Suppose F is \(\sigma_F\)-strongly convex and \(L_p\)-smooth, \(G = \omega_G\)-restricted weakly convex and \(A\) is positive definite. Then for \(0 < \alpha \leq 2\), the sequence \(\{u^k\}_{k \in \mathbb{N}}\), where \(w^k := (p^k, Bq^k, \nu^k)\) denotes the collective point at step \(k\), obtained from Solver I is bounded if \(c > \max\{\omega_G, (L_p + \sigma_F)/(\alpha \nu A^2)\}\) where \(\nu A\) is the smallest eigenvalue of a matrix \(A\). Moreover, the sequence converges to a stationary point \((p^*, Bq^*, \nu^*)\) at a linear rate locally.

Proof Sketch: The full convergence analysis is deferred to Appendix A while the details for deriving this theorem are deferred to Appendix N. We explain the key ideas in the following.

To prove convergence, the key step is to develop a sufficient decrease lemma (Lemma 6) where the difference of the objective function \(\mathcal{L}_c\) between consecutive updates (from step \(k\) to \(k + 1\)) is lower-bounded by a combination of strictly positive squared norms.

In developing Lemma 6, we expand the consecutive steps according to Solver I:

\[
\begin{align*}
\mathcal{L}_c(p^k, q^k, \nu^k) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu^{k+1}) &= \mathcal{L}_c(p^k, q^k, \nu^k) - \mathcal{L}_c(p^k, q^k, \nu_{1/2}^{k+1}) \\
&= \mathcal{L}_c(p^k, q^k, \nu_{1/2}^{k+1}) - \mathcal{L}_c(p^{k+1}, q^k, \nu_{1/2}^{k+1}) \\
&+ \mathcal{L}_c(p^{k+1}, q^k, \nu_{1/2}^{k+1}) - \mathcal{L}_c(p^{k+1}, q^k, \nu^{k+1}) \\
&+ \mathcal{L}_c(p^{k+1}, q^k, \nu^{k+1}) - \mathcal{L}_c(p^{k+1}, q^k, \nu_{1/2}^{k}) \\
&+ \mathcal{L}_c(p^{k+1}, q^k, \nu_{1/2}^{k}) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu_{1/2}^{k+1}).
\end{align*}
\]

For (12a) and (12c), due to the identity \(\|\|u\|\|u\|^2 = (1 - \eta)\|u\|^2 + \eta \|v\|^2\), we get:

\[
\begin{align*}
-\frac{c}{\alpha}\|Ap^{k+1} - Bq^{k+1}\|^2 &= -\frac{c}{\alpha}\|Ap^{k+1} - Bq^k\|^2 \\
&\leq -\frac{1}{\alpha c}\|\nu^k - \nu_{1/2}^{k+1}\|^2 - c(1 - \frac{1}{\alpha})\|\nu^k - \nu_{1/2}^{k+1}\|^2.
\end{align*}
\]

Then in (12b), since the sub-objective \(F\) is \(\sigma_F\) strongly convex and \(L_p\)-smooth, we can use Lemma 4 to find a lower bound that consists of \(\|\nabla F(p^k) - \nabla F(p^{k+1})\|^2\) and \(\|p^k - p^{k+1}\|^2\). The term with the gradient difference of \(F\) connects...
to the dual variable as \( \|\nu^k - \nu^{k+1}\|^2 = \|A^{-T}(\nabla F(p^k) - \nabla F(p^{k+1}))\|^2 \) through the first-order minimizer condition (38) and that \( A \) being positive definite. This balances the first negative squared norm in the r.h.s. of (13).

As for (12d), since \( G \) is \( \omega_G \)-restricted weakly convex (Definition 2) w.r.t. \( B \), we have a lower bound with an additional negative squared norm \( -\omega_G \|Bq^k - Bq^{k+1}\|^2 \). This negative term is balanced by the penalty coefficient \( c \) as we end up getting \((c - \omega_G)/2\|Bq^k - Bq^{k+1}\|^2\). After rearranging the terms we get the sufficient decrease lemma:

\[
\mathcal{L}_c^k - \mathcal{L}_c^{k+1} \geq \frac{\delta_p}{2}\|p^k - p^{k+1}\|^2 + \frac{\delta_q}{2}\|Bq^k - Bq^{k+1}\|^2 + \frac{\delta_\nu}{2}\|\nu^k - \nu^{k+1}\|^2,
\]

where \( \mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k) \) denotes the function value evaluated at step \( k \), and

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

From the above, we hence know the conditions in terms of the penalty coefficient \( c \) and relaxation parameter \( \alpha \) such that \( \delta_p, \delta_q, \delta_\nu \) are non-negative. Under these conditions, the sufficient decrease implies convergence. Moreover, using the sufficient decrease lemma, we can further show that the rate of convergence is locally linear.

To prove linear rate, we first show that around a local stationary point, the KL property (Definition 7) [25], [42], [43] is satisfied with an exponent \( \theta = 1/2 \), which gives:

\[
\|\mathcal{L}_c^{k+1} - \mathcal{L}_c^*\|^2 \leq K_1\|\nabla \mathcal{L}_c^{k+1}\|^2,
\]

where \( K_1 > 0 \) is a constant. The second relation that is required is as follows:

\[
\|\nabla \mathcal{L}_c^{k+1}\| \leq K_2\|w^{k+1} - w^k\|,
\]

where \( K_2 > 0 \) is another constant. If (14), (15), and the sufficient decrease lemma hold, then by Lemma 11, owing to [24], \( \theta = 1/2 \) corresponds to the desired result, i.e. the rate of convergence is \( Q \)-linear (Appendix J).

In proving (14), we separate the goal into two steps. First we find an upper bound of \( \mathcal{L}_c^{k+1} - \mathcal{L}_c^* \) (Lemma 14) by exploiting the strong convexity of \( F \) and the restricted weak convexity of \( G \):

\[
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \langle \nabla F(p^{k+1}), p^{k+1} - p^* \rangle + \langle \nabla G(q^{k+1}), q^{k+1} - q^* \rangle + \frac{\sigma_F L_p}{L_p + \sigma_F}\|p^{k+1} - p^*\|^2 + \frac{\omega_G}{2}\|q^{k+1} - q^*\|^2 + \left(\frac{\mu_A}{2}\right)^2\|A p^{k+1} - B q^{k+1}\|^2.
\]

Then by the first-order necessary conditions (38) for the gradients of \( F, G \), along with the relation \( Ap^* = Bq^* \) at a stationary point, we obtain the following:

\[
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \frac{C}{2}\|A p^{k+1} - A p^*\|^2 + \frac{\omega_G - C}{2}\|B q^{k+1} - B q^*\|^2.
\]

Second, using (38) again along with the assumption that the matrix \( A \) is positive definite, we get:

\[
\|\nabla \mathcal{L}_c^{k+1}\|^2 \geq K_2\|A p^{k+1} - B q^{k+1}\|^2.
\]

Combining the two inequalities through the Cauchy-Schwarz inequality \( \|u - v\|^2 \geq (1 - t)\|u\|^2 + (1 + 1/t)\|v\|^2, t > 0 \), we prove that the KL exponent \( \theta = 1/2 \). As for (15), following similar reasoning for (17), we have:

\[
\|\nabla \mathcal{L}_c^{k+1}\|^2 \leq K_3\|A p^{k+1} - B q^{k+1}\|^2 \\
\leq 2K_3\|A p^{k+1} - B q^k\|^2 + \|B q^{k+1} - B q^k\|^2,
\]

where \( K_3 > 0 \). Substitute (13) into the first term in the r.h.s. of the above inequality, assuming \( 1 \leq \alpha \leq 2 \), we arrive at the desired result. Note that the relaxation parameter we focus on has range \( 0 < \alpha < 2 \), and the above \( Q \)-linear rate of convergence result does not apply to \( 0 < \alpha < 1 \), to address this, inspired by the recent work [21], we prove the \( R \)-linear rate of convergence for this region. The key relation that enables this result is:

\[
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq K_4 (\mathcal{L}_c^{k+1} - \mathcal{L}_c^* + \|w^{k+1} - w^*\|^2),
\]

where \( K_4 > 0 \). It turns out that if the sequence of the function value \( \{\mathcal{L}_c^k\}_{k \in \mathbb{N}} \) converges \( Q \)-linearly, implied by the sufficient decrease lemma, and (18) holds, then we can prove \( R \)-linear rate of convergence of the sequence \( \{w_k\}_{k \in \mathbb{N}_0}, k > N_0 \in \mathbb{N} \) locally around the neighborhood of a stationary point \( w^* \) (Lemma 13, proved in Appendix K).

In proving (18), from the sufficient decrease lemma, we can substitute \( c > \omega_G \) into (16) and get:

\[
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \frac{c\mu_A^2}{2}\|p^{k+1} - p^*\|^2 \\
\leq \frac{c\mu_A^2}{2}\|w^{k+1} - w^*\|^2 + \rho^*\|w^{k+1} - w^*\|^2.
\]

Then the desired relation can be obtained by letting \( K_4 := \max\{c\mu_A^2/2, \rho^*\} \).

Note that Theorem 1 requires the function \( F \) to be strongly convex and \( G \) to be restricted weakly convex which limits the class of functions that the theorem applies to. To address this, we adopt Solver II whose convergence is also locally linear but only requires \( F \) to be convex and \( G \) to be weakly convex.

**Theorem 2:** Let \( F \) be convex, \( G \) be \( \sigma_G \)-weakly convex and \( L_q \)-smooth, \( A \) be full row rank and \( B \) be positive definite, then for \( 0 < \alpha < 2 \) the sequence \( \{w_k\}_{k \in \mathbb{N}} \) where \( w_k := (A p^k, q^k, \nu^k) \), obtained from Solver II is bounded if

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

Moreover, the sequence converges to a stationary point \( w^* := (A p^*, q^*, \nu^*) \) on a linear rate locally.

**Proof Sketch:** The details are deferred to Appendix Q. The steps are similar to the proof sketch for Theorem 1. For the sufficient decrease lemma (Appendix G), we divide \( \mathcal{L}_c^{k} - \mathcal{L}_c^{k+1} \)
according to (11). For the relaxation and dual steps, using identity (36), we have:

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

(19)

As for the $p$-update, by the convexity of $F$, first-order necessary conditions (40), and the identity $2(u - v, w - u) = \|v - w\|^2 - \|u - v\|^2 - \|u - w\|^2$, we have:

$$L_c(p^k, q^k, \nu^k) - L_c(p^{k+1}, q^k, \nu^k)$$

$$\geq \langle \nabla F(p^{k+1}), p^k - p^{k+1} + \langle \nu^k, Ap^k - Ap^{k+1} \rangle$$

$$+ \frac{c}{2}\|Ap^k - Bq^k\|^2 - \frac{c}{2}\|Ap^{k+1} - Bq^k\|^2$$

$$\geq \frac{c}{2}\|Ap^k - Ap^{k+1}\|^2.$$

Lastly, for the $q$-update, by $\sigma_G$-weak convexity of $G$:

$$L_c(q^k) - L_c(q^{k+1}) \geq -\frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 + \frac{c}{2}\|Bq^{k+1} - Bq^k\|^2.$$

To balance the negative squared norm, note that in $Solve II$, the $q$-update precedes the dual update, this connects the gradient of $G$ to the dual variable $v$ as: $\nabla G(q^{k+1}) = B^T \nu^{k+1}$. This relation along with the assumption that $G$ is $L_q$-smooth and $B$ is positive definite results in the following:

$$\|\nu^k - \nu^{k+1}\|^2 \leq \mu_B^2\|\nabla G(q^k) - \nabla G(q^{k+1})\|^2$$

$$\leq \frac{L_i^2\mu_B^2}{2}\|q^k - q^{k+1}\|^2.$$

(20)

With (20) and combining the above results, we obtain:

$$L_c^k - L_c^{k+1} \geq \left\{ \begin{array}{ll}
\frac{c\mu_B^2}{2} \left[ \frac{1}{\alpha} - \frac{1}{2} \frac{L_q^2}{\sigma_G} \right] - \frac{\sigma_G}{2} \|q^k - q^{k+1}\|^2.
\end{array} \right.$$

(21)

Then it is straightforward to determine the range in terms of the penalty coefficient $c$ such that the coefficient of $\|q^k - q^{k+1}\|^2$ is positive. This proves the sufficient decrease lemma (Lemma 8). Then we proceed to prove the linear rate of convergence. Similar to the proof sketch for Theorem 1, given the sufficient decrease lemma, the next step is to show that the inequalities (14) and (15) hold. From the minimizer conditions and the assumption that $A$ is full row rank, we develop Lemma 17, which gives:

$$\|\nabla L_c^{k+1}\| \geq K_5(\|\nu^k - \nu^{k+1}\|^2 + \|q^k - q^{k+1}\|^2),$$

and

$$L_c^{k+1} - L_c^* \leq K_6(\|\nu^k - \nu^{k+1}\|^2 + \|q^k - q^{k+1}\|^2 + \|w^{k+1} - w^*\|^2),$$

where $K_5, K_6 > 0$. Combining the above two inequalities:

$$L_c^{k+1} - L_c^* \leq K_7(\|\nabla L_c^{k+1}\|^2 + \|w^{k+1} - w^*\|^2)$$

$$\leq K_7\|\nabla L_c^{k+1}\|^2,$$

where $K_7 > 0$ and in the last inequality we consider a neighborhood around $w^*$ such that $\|w^{k+1} - w^*\| < \varepsilon, L_c^* < L_c^{k+1} < L_c^* + \delta$, and use Lemma 12. This assures the existence of $\|\nabla L_c^{k+1}\| > \eta$ for $\eta > 0$ if $w^{k+1} \notin \Omega^*$. These results proves that the Łojasiewicz exponent $\theta = 1/2$ locally around $w^*$ and that (14) is satisfied. As for (15), since $A$ is full row rank, from the first-order necessary conditions (40), we have:

$$\|\nabla L_c^{k+1}\|^2 \leq K_8(\|w^k - w^{k+1}\|^2 + \|q^k - q^{k+1}\|^2 + \|Ap^{k+1} - Bq^k\|^2),$$

where $K_8 > 0$. For $1 \leq \alpha < 2$, we can find an upper bound for $\|Ap^{k+1} - Bq^k\|^2$ through (19), and obtain:

$$\|\nabla L_c^{k+1}\| \leq K_9\|w^{k+1} - w^*\|,$$

where $K_9 > 0$. This proves the relation (15). Consequently, by Lemma 11 the rate of convergence for $1 \leq \alpha < 2$ is $Q$-linearly. Finally, for the region $0 < \alpha < 1$, from (22), (20) and the sufficient decrease lemma:

$$L_c^{k+1} - L_c^* \leq K_{10}(L_c^* - L_c^{k+1}) + \|w^{k+1} - w^*\|^2,$$

where $K_{10} > 0$. This shows that the relation (18) holds, and by Lemma 13, we conclude that $\{w^k\}_{k>N_0}, N_0 \in \mathbb{N}$ converges $R$-linearly. Together, we prove that the rate of convergence is locally linear for $0 < \alpha < 2$.

Compared to Theorem 1, Theorem 2 needs the matrix $B$ to be positive definite instead of $A$. The change is necessary to balance the weak convexity of the sub-objective $G$. In the Markovian optimization problem (6) we considered, both the two theorems apply. This is because the linear constraints in fact represent the marginal or Markov relations between the (conditional) probabilities. Therefore, one of the matrices $A$ and $B$ is an identity matrix whereas the other is full row rank. Inspired by the recent results [37], we develop the following theorem that keeps the definiteness of the matrix $A$ as in Theorem 1 with relaxed properties as in Theorem 2, but require $G$ to have Lipschitz continuity additionally. The extra assumption allows us to have the reverse norm bound:

$$\|q^m - q^m\| \leq M_q\|Bq^m - Bq^m\| \text{ without $B$ being positive definite, balancing the weak convexity of $G$.}$$

**Theorem 3:** Suppose $F$ is convex and $L_y$-smooth, $G$ is $\sigma_G$-weakly convex, $L_q$-smooth, and $M_q$-Lipschitz continuous, and the matrix $A$ is positive definite while $B$ is full row rank, then for $0 < \alpha < 2$, the sequence $\{w^k\}_{k \in \mathbb{N}}$, where $w^k := (p^k, q^k, \nu^k)$, obtained from Solver II is bounded if:

$$c > M_q \left[ \frac{M_q \sigma_G \alpha + \sqrt{M^2_q \sigma^2_G \alpha^2 + 8(2 - \alpha) \lambda_B^2 \mu_B^2}}{4 - 2\alpha} \right],$$

where $\lambda_B$ denotes the largest positive singular value of the matrix $B$; $\mu_B^T$ the smallest eigenvalue of the matrix $B^T$. Moreover, the sequence converges to a stationary point $w^* := (p^*, q^*, \nu^*)$ at a linear rate locally.

**Proof Sketch:** The details are deferred to Appendix S. The only difference as compared to the proof sketch of Theorem 2 lies in the regularities of the linear constraints, where now $A$ is positive definite and $B$ is full row rank. For establishing the sufficient decrease lemma, and hence the convergence (Lemma 10), we exploit the Lipschitz continuity of $G$. Since the $q$-update can be equivalently expressed as a function $\Psi(u) := \arg\min_{q \in \Omega_q} G(q) + c/2\|Bq - u\|^2$,
as in [37], we have \( \|q^m - q^n\| = \|\Psi(Bq^m) - \Psi(Bq^n)\| \leq M_q \|Bq^m - Bq^n\| \). Replacing the terms \( \|Bq^k - Bq^{k+1}\|^2 \), or \( \|Bq^{k+1} - Bq^k\| \) with this relation, we can prove the local linear rate for the region of the relaxation parameters \( 0 < \alpha < 2 \).

As a remark, while we focus on entropy and conditional entropy functions for applications, the convergence analysis for the algorithms hold for general functions satisfying the assumptions mentioned in the theorems. This is closely related to the recent strongly-weakly-convex pair problems [27], [36]. This class of non-convex functions are less explored until recently in contrast to the well-studied convex counterpart. For the convexity, the negative as in [37], we have

\[
\inf_{x \in X} u(x) = \epsilon.
\]

The \( \epsilon \)-infinite assumptions are commonly adopted in density/entropy estimation problems [29], [30] for smoothness conditions that will facilitate the optimization process. A key result used in this work is that the entropy function whose associated probability mass vector is an \( \epsilon \)-infinite measure, is weakly convex with a coefficient proportional to \( 1/\epsilon \). In addition, under the Markov chain \( Y \rightarrow X \rightarrow Z \), by data-processing inequality we can have a tighter bound than the weak convexity, which is defined as the restricted weak convexity.

Definition 2: A function \( f : \mathbb{R}^d \rightarrow [0, \infty) \), is \( \omega \)-restricted weakly convex, \( \omega > 0 \) w.r.t. a matrix \( A \in \mathbb{R}^{m \times d} \) if \( f \in C^1 \) and the following holds:

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle - \omega \|Ay - Ax\|^2.
\]

It is worth noting that the restricted weak convexity was adopted in deriving the privacy parameter in differential privacy [44]. As will be shown below, it relaxes the regularity conditions of the linear constraints.

**IV. Applications**

In the section, we apply the main results stated above to practical problems. In the general Markovian Lagrangian framework (6), we focus on two specific non-convex information theoretic optimization problems, i.e., the IB and PF problems.

When applied to IB and PF, we consider vector variables whose elements are composed of the vectorized, discrete (conditional) probability mass, defined as follows:

\[
p_{z|x} := \begin{bmatrix} p(z_1|x_1) \cdots p(z_N|x_N) \\ \vdots \\ p(z_N|x_N) \end{bmatrix},
p_{z} := \begin{bmatrix} p(z_1) \cdots p(z_N) \end{bmatrix}^T,
p_{z|y} := \begin{bmatrix} p(z_1|y_1) \cdots p(z_N|y_N) \\ \vdots \\ p(z_N|y_N) \end{bmatrix},
\]

where \( N := |\mathcal{V}| \) denotes the cardinality of a variable \( \mathcal{V} \). In both IB and PF, the variable to optimize is the conditional probability \( p_{z|x} \). Therefore, one of the variables \( p, q \) in (9) must be \( p_{z|x} \). As for the other one, it can be assigned as \( p_z, p_{z|y} \) or formed by cascading \( p_z \) and \( p_{z|y} \). The relation between the conditional and marginal probabilities becomes a linear penalty \( AP - Bq \) with each row of \( A, B \) being a prior probability vector. For example, if we let \( p := p_z \) and \( q := p_{z|x} \), then \( A := I, B := I \otimes p_z^T \) where \( \otimes \) denotes the Kronecker product. Under this construction, the linear penalty term penalizes the case where the marginal probability relation \( \sum_x p(z|x)p(x) = p(z) \) is violated.

Our results depend on the convexity and smoothness of the two sub-objective functions. For the convexity, the negative entropy function \(-H(X)\) is convex w.r.t. the probability \( p_z \). Similarly, for the negative conditional entropy with a known marginal, it is also convex [28]. Additionally, we use the following definition to establish smoothness conditions for the (conditional) entropy functions and find the associated Lipschitz coefficients.

**Definition 1:** A measure \( u(x) \) is said to be \( \epsilon \)-infinite if there exists \( \epsilon > 0 \), such that \( \inf_{x \in X} u(x) = \epsilon \).

As shown in Table I, our results are based on the (strongly convex-weakly convex) structure of the two sub-objective functions. Interestingly, the IB problem satisfies these conditions. To see this, we use the known result in IB that when the trade-off parameter \( \gamma \geq 1 \), the corresponding minimum loss for the IB Lagrangian (2) are trivial, e.g., \( I(Z;X) = I(Z;Y) = 0 \) [45]. Based on this result, we can focus on the region \( 0 < \gamma < 1 \) accordingly, hence \( (\gamma - 1)H(Z) \) is strongly convex w.r.t. \( p_z \) and \(-\gamma H(Z|X) \) is convex w.r.t. \( p_{z|x} \). Recall that for the IB problem, the associated coefficient set in (6) is \( p_z = \gamma - 1, p_{z|x} = -\gamma \) and \( p_{z|y} = 1 \), so there exist multiple combinations in terms of the variables \( p_z, p_{z|x}, p_{z|y} \) to apply our framework. In particular, we propose two splitting methods for the IB problem that work on **Solver I** and **Solver II** respectively. By imposing \( p_{z|y} = Q_{z|y}p_{z|x}^k \) as an equality constraint, we can have the following re-formulation of the IB problem to the proposed framework (9) as:

\[
p := p_z,
q := p_{z|x},
p_{z|y} = Q_{z|y}p_{z|x},
F(p) := \langle \gamma - 1 \rangle H(Z),
G(q) = -\gamma H(Z|X) + H(Z|Y),
A = I_N, 
B = Q_x.
\]

where the matrix \( Q_{z|y} \) is defined to express the Markov relation \( p(z|y) = \sum_x p(z|x)p(x|y) \) as a linear equation. Specifically, if we represent the conditional probability \( p(x|y) \) as a matrix \( W_{z|y} \in \mathbb{R}^{N_z \times N_y} \) with the \((i,j)\)-entry \( w_{ij} = p(x_i|y_j) \), then \( Q_{z|y} := I_{N_z} \otimes W_{z|y} \) where \( \otimes \) denotes the Kronecker product. Similarly, we have \( Q_x := I_{N_x} \otimes p_z^T \) that maintains the marginal relation between \( p_z \) and \( p_{z|x} \). From above, we can treat the Markovian relation \( p_{z|x} = Q_{z|x}p_{z|x} \) as an additional penalty which gives an alternative form to (23). This corresponds to the following settings:

\[
p := p_{z|x},
q := \begin{bmatrix} p_z^T \\ p_{z|y}^T \end{bmatrix}^T.
\]
\[ F(p) := -\gamma H(Z|X), \]
\[ G(q) := (\gamma - 1)H(Z) + H(Z|Y), \]
\[ A = \begin{bmatrix} Q_x^T & Q_{z|y}^T \end{bmatrix}^T, \quad B = I_{N_q}, \quad (24) \]

where \( N_q := |Z| \times (|Y| + 1) \). Interestingly, we find that the first formulation (23) satisfies the assumptions of Theorem 1 while the second one (24) meets the assumptions of Theorem 2. Therefore, we have the following convergence guarantees for the two proposed IB solvers.

**Theorem 4:** Suppose \( p_z \) is \( \varepsilon_z \)-inifimal and \( p_{z|x} \) is \( \varepsilon_{z|x} \)-inifimal, then for \( 0 < \alpha \leq 2 \), the IB problem formulated as in (23) satisfies:

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

Moreover, let \( w^k := (p_{z|x}, Q_{z|y}, p_{z|x}, \nu^k) \), then the sequence \( \{w^k\}_{k \in \mathbb{N}} \) converges at a linear rate to a stationary point locally when solved with Solver I with a penalty coefficient as required in Theorem 1.

**Proof:** See Appendix T.

The second IB solver corresponds to the construction as in (24), whose convergence guarantee is based on Theorem 2.

**Theorem 5:** Suppose \( p_z, p_{z|y}, p_{z|x} \) are \( \varepsilon_z, \varepsilon_{z|y}, \varepsilon_{z|x} \)-inifimal, respectively, then for \( 0 < \alpha < 2 \), the IB problem formulated in (24) satisfies:

- \( F(p) \) is convex and \( 1/\varepsilon_{z|x} \)-smooth,
- \( G(q) \) is \( \sigma_G \)-weakly convex and \( L_q \)-smooth,
- The matrix \( A := \begin{bmatrix} Q_x^T & Q_{z|y}^T \end{bmatrix} \) is full row rank.

Moreover, the sequence \( \{w^k\}_{k \in \mathbb{N}} \) converges at a linear rate to a stationary point locally when using Solver II with a penalty coefficient satisfying the requirements in Theorem 2, where \( \sigma_G := (2N_zN_y)/\varepsilon_{z|y} \) and \( L_q = \max\{1/\varepsilon_z, 1/\varepsilon_{z|x}\} \).

**Proof:** See Appendix U.

Notably, the first form (23) is proposed in our earlier work [16], but it is limited to \( \alpha = 1 \). While the convergence is proved therein, an additional Bregman divergence is added to the \( p \)-update to regularize it, and to facilitate convergence analysis. In contrast, using Theorem 4 we show that the convergence can be proved without additional regularization. On the other hand, the second solver is new in splitting methods for IB to our knowledge.

Theorem 4 and Theorem 5 allow us to compare the two DRS algorithms in terms of the smallest penalty coefficient that assures convergence.

From the formulation and convergence analysis, the advantages of each solver are clear. For Solver I, it has fewer augmented variables to optimize since it is restricted to solutions where \( p_{z|y} = Q_{z|y}p_{z|x} \) holds. On the other hand, for Solver II, the smallest penalty coefficient that assures convergence \( c_{IF}^\alpha \) is independent of \( L_q \) and \( \sigma_F \), except for the case where \( p(z|y) \) is deterministic. The independence is useful in evaluating the solutions for the IB problem on the information plane [9], [46]. A common practice to form the information plane is to vary the trade-off parameter \( \gamma \) over a certain range. Since varying \( \gamma \) does not change the weak-convexity coefficient of

**Solver II.** This invariance therefore allows us to fix a penalty coefficient \( c_H \) when collecting IB solutions from varying \( \gamma \).

We further compare the convergence rates of the two proposed IB solvers to existing ones. By Theorem 4 and 5, the convergence rates are locally linear. Remarkably, the BA-based algorithm, often serves as a benchmark, is also known to be convergent with linear rate [47]. Moreover, empirically the two solvers can obtain solutions with tantamount performance on the information plane.

Remarkably, our theoretic results can extend further than solving the IB Lagrangian. Inspired by the variational inference method on IB [2], we apply our solvers to a surrogate upper bound of the IB Lagrangian:

\[
F(p) := -\gamma H(Z|X) - \sum_{z \in Z} \sum_{y \in Y} p(z, y) \log q(y|z),
\]
\[
G(q) := \gamma H(Z),
\]
\[
A = Q_{z|y}, \quad B = I_{N_q}, \quad (25)
\]

where \( q(y|z) \) is the variational parameter and is updated as in (4c) after an iteration of Solver II. Observe that the construction satisfies the requirements in Theorem 5, with \( p_{z|x} \) viewed as a function of \( p_{z|y} \) and the weakly convex sub-objective reduces to \( \gamma H(Z) \) only, hence the convergence guarantee also applies to this variant solver.

**B. Solvers for Privacy Funnel**

Our general framework includes the PF problem as a special case as well. We can decompose the PF problem into a convex \(-\beta H(Z|Y)\) w.r.t. \( p_{z|y} \) and weakly convex \((\beta - 1)H(Z) + H(Z|X)\) w.r.t. \( p_{z|x} \) pair of objectives. This is due to Lemma 2 under \( \varepsilon_{z|x} \)-inifinality and that \( p_z = Q_xp_{z|x} \). The decomposition allows us to map the PF problem to the augmented Lagrangian as follows:

\[
p := p_{z|y}, \quad q := p_{z|x}, \quad p = Q_xp_{z|x}, \quad F(p) := -\beta H(Z|Y), \quad G(q) := (\beta - 1)H(Z) + H(Z|X), \quad A = I_{N_z}, \quad B = Q_z|y. \quad (26)
\]

The weak convexity coefficient is determined through Lemma 5 with \( p_{z|x} \) assumed to be \( \varepsilon_{z|x} \)-inifimal. Under this construction, we can solve the PF problem with Solver II. Moreover, we have the following convergence guarantee for the new PF solver.

**Theorem 6:** Suppose \( p_{z|y}, p_{z|x} \) are \( \varepsilon_{z|y}, \varepsilon_{z|x} \)-inifimal respectively, then for \( 0 < \alpha < 2 \), the PF problem formulated in (26) satisfies:

- \( F(p) \) is convex and \( 1/\varepsilon_{z|x} \)-smooth,
- \( G(q) \) is \( \sigma_G \)-weakly convex, \( L_q \)-smooth and \( M_q \)-Lipschitz continuous.
- The matrix \( B := Q_z|y \) is full row rank.

Moreover, let \( w^k := (p_{z|y}^k, p_{z|x}^k, \nu^k) \), then the sequence \( \{w^k\}_{k \in \mathbb{N}} \) converges at linear rate locally to a stationary point when using Solver II with a penalty coefficient satisfying the requirements in Theorem 3, where the coefficients...
\[ \sigma_G := \frac{1}{2N_x}(|\beta - 1| + N_y) / \varepsilon_{z|x}, L_q := 1 / \varepsilon_{z|x}, M_q := 2 \log \varepsilon_{z|x} \] and \( \mu_{BB}^2 \) is the smallest eigenvalue of the matrix \( BB^T \).

Proof: See Appendix V.

In literature, most PF solvers satisfying our Markovian framework are based on the agglomerative clustering approach [9] which is restricted to deterministic mappings [3], [8]. In contrast to these works, the new proposed PF solver can recover solutions obtained by the clustering based PF solvers on the information plane. Moreover, as shown in our numerical results, we can achieve lower privacy leakage compared to them.

Similar to applying the proposed methods to variational inference-based IB, we can solve a surrogate upper bound of the PF Lagrangian (3), which is provided as follows:

\[
\mathcal{L}_{V1,PF} := \beta H(Z) - \beta H(Z|Y) - \sum_{z \in Z} \sum_{x \in \mathcal{X}} p(z, x) \log q(x|z),
\]

where \( q^{k+1}(x|z) = \frac{p(x)p^k(z|x)}{\sum_x[p(x)p^k(z|x)]} \) denotes the variational distribution, satisfying the requirements stated in Theorem 5:

\[
p := p_{z|x}, q := p_{z|x}, p_{z|y} = Q_{x|y}p_{z|x}, F(p) := -\beta H(Z|Y) - \sum_{z \in Z} \sum_{x \in \mathcal{X}} p(z, x) \log q(x|z),
\]

\[
G(q) := \beta H(Z),
\]

\[
A = Q_{x|y}, B = I_{N_z},
\]

Note that \( F \) is a convex function w.r.t. \( p = p_{z|x} \) because the variational decoder \( q(x|z) \) is fixed during \( p, q \) updates. In contrast to existing variational PF solvers whose decoder \( q(x|y, z) \) depends on the sensitive information \( Y \) [5], the variational decoder in (27) can be optimized without passing \( Y \) as a required input to the decoder.

V. Evaluation

In this section, we present simulation results for the proposed algorithms using synthetic and real-world datasets. We implement the solvers with gradient descent to update the variables \( p_z, p_{z|x}, p_{z|y} \). To ensure that the updated variables remain valid probability vectors, projection is needed [41]. There are various ways to project the updated variables to a probability simplex, the one we implemented is known as the mean-subtracted gradient because we empirically find that this method is more efficient for the linearly-constrained augmented Lagrangian (6). The mean-subtracted gradient is given by:

\[
p^{k+1} = p^k - \xi_k \nabla \hat{E}_c, \quad \nabla \hat{E}_c = \nabla E_c - \frac{1}{N_p} \sum_{i=1}^{N_p} \nabla E_{c,i},
\]

where \( \xi_k \) is a sufficiently small step-size at step \( k \). This method introduces an extra parameter, the step-size to decide, where the standard back-tracking line-search is adopted [41].

All the proposed solvers are initialized as follows, we use Python Numpy package to sample from a Unif(0,1) source randomly for \( |Z| \times |X| \) times and arrange them into a \( R^{(|Z| \times |X|)} \) matrix. Then the entries are normalized. The main focus of the evaluation is the characterization of the information plane of either an IB or PF solver.

A. Datasets

The synthetic conditional distribution used in our evaluation is given by:

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

Additionally, we evaluate the performance with the following non-uniform \( p(X) \).

\[
p_{non-unif}(X) = \begin{bmatrix} 0.1 & 0.3 & 0.6 \end{bmatrix}^T.
\]

We set the representation dimension \( |Z| \leq 4 \) for both the IB and PF [48].

We also evaluate the proposed methods on a real-world dataset. The dataset is named “Heart failure clinical records Data Set” [49] from the UCI Machine Learning Repository [50]. 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. We let \( Y := \{ \text{"sex"}, \text{"death"} \} \) and the rest be \( X \); this results in \(|Y| = 4, |X| = 16\). The joint probability is formed by counting the 299 instances w.r.t. \( (Y, X) \) pair. The counted results are post-processed by adding \( 10^{-3} \) to each entry to avoid \( p(x, y) = 0 \). Finally, the cardinality of the representation satisfies \(|Z| \leq 17\) due to [48].

B. Privacy Funnel

The proposed PF solver is denoted as Solver II, consistent with our earlier notation. The corresponding variational inference variant of this solver (28) is denoted as Solver II-V. We initialize the encoder \( p(z|x) \) as described in Section V, obtaining a feasible point in the compound simplex \( \Omega_{z|x} \). Since \( p(X, Y) \) is assumed to be known, computing both \( p(z) = \sum_x p(z|x)p(x) \) and \( p(y) = \sum_y p(z|y)p(x|y) \) are straightforward.

For the synthetic dataset, the range of the trade-off parameters is set as \( \beta \in [0.5, 20.0] \) and we generate 20 geometrically-spaced grid points of the range. For a \( \beta \) within this range, 10 trails are performed where the termination condition is either: 1) convergent when total variation is lower than a pre-determined small threshold, e.g., set to \( \|Ap - Bq\| \lesssim 2 \times 10^{-6} \) or 2) divergent when a maximum number of iterations is reached otherwise. In the convergent case, the resulting encoder probability \( p_{z|x} \) can be used to compute the mutual information pair \( I(Z; X), I(Z; Y) \). We collect convergent cases, calculate the resultant mutual information pairs \( I(Z; X), I(Z; Y) \) and then use them to characterize the privacy-utility trade-off, i.e., find the lowest \( I(Z; Y) \) for a fixed \( I(Z; X) \).

In Fig. 1a, we compare the information plane of the two proposed solvers on the synthetic dataset with the uniformly

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 1. Information planes of PF solvers using the synthetic dataset with a uniform $p(X)$. Fig. 1a compares the two proposed PF solvers whereas Fig. 1b, compares the best proposed solver with the clustering-based PF solver.

Fig. 2. Information planes of PF Solvers with a non-uniform $p_{\text{non-unif}}(X)$. Fig. 2a compares the two proposed PF solvers. In Fig. 2b the best proposed solver is compared to the clustering-based PF algorithm.

distributed marginal $p_{\text{unif}}(X)$. The first solver is denoted as Solver II and the relaxation parameter is set to $\alpha = 1.618$. This solver is based on Theorem 6. The second proposed solver is the variational inference-based solver (27) where a surrogate upper bound to the PF Lagrangian (3) is minimized through Solver II with $\alpha = 1.000$. We denote the method as Solver II-V. As shown in Fig. 1a, under the same range of the trade-off parameter $\beta$ and the same number of trials, the Solver II performs better (i.e., achieves a lower information leakage). In Fig. 2a, we further consider the non-uniform marginal $p_{\text{non-unif}}(X)$ where it is shown that Solver II-V is better instead (refer to the $I(Z;Y)$ in the range $I(Z;X) \in [0.8, 1.2]$ bits).

Then we evaluate the proposed solvers on the real-world dataset. In this experiment, the trade-off parameter has a range $\beta \in [1.0, 10.0]$ and 20 trials are performed. The results are shown in Fig. 3. In Fig. 3a, we compare the two proposed methods. For Solver II the penalty coefficient is tuned to 7000 while for the Solver II-V the penalty coefficient is tuned to 128. In this experiment setup, Solver II is found to perform better. Note that both solvers achieve the “perfect privacy” [6] region (i.e., the utility $I(Z;X) > 0$ while $I(Z;Y) \approx 0$).

We further examine the convergence behavior of Solver II. Recall that our theoretical results imply that the penalty coefficient needs to be larger than a threshold to ensure convergence. In the real-world dataset, we fix a trade-off parameter $\beta = 3.5$ and sweep through a range of the penalty coefficient $c \in [1000, 8000]$. For each $c$, we perform 2000 trials and calculate the percentage of convergent cases. The results are shown in Fig. 4a, where we compare the two cases $\alpha = 1.618$ and $\alpha = 2.000$. We observe that to achieve 80% of convergent cases, the smallest $c$ for $\alpha = 2.000$ is lower than that of $\alpha = 1.618$, which aligns with our theoretical predictions. Lastly, we examine the rate of convergence in Fig. 4b. We compare two cases $\alpha = 1.618$ and $\alpha = 2.000$, fixing the penalty coefficients to $c = 7000$. The minimum loss is $\hat{L}^*_{c} = -2.46$. The loss decrease of Solver II shown in Fig. 4b clearly reflects a locally linear rate. Therefore this result is numerical evidence supporting our theory.

In Fig. 1b, we compare our new solver with the state-of-the-art clustering based algorithm (referred to as Merge-Two) under the synthetic dataset with $p_{\text{unif}}(X)$. We observe that in the range $I(Z;X) \in [0, 0.9]$ bits, our solver obtains more points on the information plane than Merge-Two, and these
Fig. 3. Proposed PF Solvers-vs-Clustering based algorithms.

points have lower privacy leakage. However, the proposed solver converges to a local minimum at $I(Z; X) \approx 0.9$. This can be improved through a more optimized implementation which is the subject of our future work. In 2a we repeat the comparison with the non-uniform $p_{\text{non-unif}}(X)$. Solver II-V is used here since it provides the best result. Again, we observe that Solver II-V recovers the solutions of Merge-Two around $I(Z; X) \approx \{0.5, 1.0, 1.3\}$ bits and achieves lower privacy leakages otherwise. Finally, Fig 3b reports the comparison with the real-world data set where the same trend is observed.

C. Information Bottleneck

We adopt the same numerical setup for the synthetic dataset with both uniform and non-uniform marginal probabilities $p_{\text{unif}}(X)$ and $p_{\text{non-unif}}(X)$. The trade-off parameter $\gamma \in [0.1, 1.0]$ and 16 geometrically-spaced grid points are evaluated. For each $\gamma$, 16 trials are performed. Each trial is initialized as described in Section V. The same convergence criterion for the proposed PF solvers is adopted here. For the information plane, only convergent cases are considered when characterizing the relevance-complexity trade-off. For the proposed solvers, we denote Solver I for (23) and Solver II for (24). The proposed variational inference-based solver in (25) is denoted as Solver I-V.

We first evaluate the proposed solvers on $p_{\text{unif}}(X)$ in Fig. 5a with $c_{I} = 16, \alpha_{I} = 1.618, c_{II} = 64, \alpha_{II} = 1.000$ and $c_{I-V} = 64$. The proposed solvers mostly obtain comparable Pareto-frontier solutions but we observe that Solver I-V converges to a local minimum at $I(Z; X) \approx 0.8$ bits whereas Solver I and Solver II do not. Then we evaluate the three solvers on the non-uniform $p_{\text{non-unif}}(X)$ in Fig. 6a. We observe that for $I(Z; X) \in [0.2, 0.6]$ bits, Solver II converges to slightly sub-optimal solutions while in $I(Z; X) \in [0.85, 1.2]$ bits Solver I-V converges to sub-optimal solutions, and hence, Solver I provides the best performance.

Next, we evaluate the convergence performance of the proposed solvers on the synthetic dataset with uniform $p_{\text{unif}}(X)$. In this simulation, the trade-off parameter is set to $\gamma = 0.20$. Each solver starts from a randomly initialized point (the same method as in the last part) for 100 trials. In Fig. 7a Solver I is evaluated with two settings $\alpha = 1.000$ and $\alpha = 2.000$. We observe that to reach 90% convergent percentage, $\alpha = 2.000$ requires a smaller penalty coefficient compared to $\alpha = 1.000$ which aligns with our theoretical convergence analysis (Theorem 4). Similar observations can be found in
Fig. 5. Information planes of IB solvers evaluated on the synthetic dataset with uniform $p_{\text{unif}}(X)$.

Fig. 6. Information planes of IB solvers evaluated on the synthetic dataset with non-uniform $p(x)$.

Fig. 7. Convergence Evaluation of the Proposed IB Solvers evaluated on the synthetic dataset with a uniform marginal $p_{\text{unif}}(X)$. In Fig. 7c we compare the loss decrease versus number of iterations.

Fig. 7b where Solver II is configured with $\alpha = 1.000$ and $\alpha = 2.000$. Clearly, the case $\alpha = 2.000$ requires a smaller $c$ for to reach the same convergence percentage. This aligns with Theorem 5.

In Fig. 7c, we compare the convergence behavior of Solver I and Solver II. We fix $\gamma = 0.20$ and initialize both solvers from the same point. We run the two solvers until convergence and compare their loss decrease. We report the fastest (lowest number of iterations) configurations of each solver. In this specific case, the minimum loss among the two solvers is $L^* = -0.324$. As shown in the figure, both solvers first explore the loss surfaces at sub-linear rates, then when the solvers operate within a neighborhood of a stationary point, they converge to it linearly. Finally, Solver I is shown to require a smaller penalty coefficient than Solver II.
In this section, we prove the convergence and the corresponding rates for Solver I and Solver II. We start with the preliminaries including definitions and properties that will be used in the proofs.

**A. Preliminaries**

**Definition 3:** A function \( f : \mathbb{R}^n \mapsto [0, \infty) \), with distinct \( x, y \in \Omega \) is Lipschitz continuous if:

\[
|f(x) - f(y)| \leq L|x - y|,
\]

where \( L > 0 \) is the Lipschitz coefficient. This motivates exploring more efficient implementation, within the same splitting methods framework, which is an interesting topic for future work.

Finally, we evaluate the proposed IB solvers on the real-world dataset. The result is shown in Fig. 9. The range of the trade-off parameter is \( \gamma \in [0.01, 1.0] \) and 64 geometrically-spaced grid points are generated from this range. For a \( \gamma \in \) in the range, 16 trails are performed by each algorithm. We collect the obtained solutions and plot the convergent cases on the information plane. The proposed method is Solver I-V as we empirically found that it performs best among the proposed solvers. Compared to BA, the proposed solver can span the information plane with more points on the Pareto-frontier (observe \( I(Z;X) \in [0, 0.5] \) bits).

**VI. Conclusion**

In this work, we considered a general discrete rate distortion Lagrangian following a three letter Markov chain. The general framework includes the IB and PF problems as special cases. We proposed solving the general problem with splitting methods that are capable of handling large-scale problems. This includes important applications in multi-view learning [51], [52] and multi-source privacy problems [7], [53], [54], [55].

Our convergence analysis is general for any objective function that can be decomposed to a convex-weakly convex pair. We further proved that our proposed solvers are linearly convergent. Based on these theoretical insights, we developed optimized new solvers for both the IB and PF problems. For the two classes of the developed IB solvers, the first class has fewer variables to optimize by restricting the Markov relation to hold strictly while the second class is convergent independent of the selection of the trade-off parameter, controlling the relevance-compression trade-off (except for one special case). In the PF case, our new solvers are shown to outperform the state of the art clustering-based solvers. Our empirical evaluations include synthetic and real-world datasets and explore both uniform and non-uniform priors.

For future work, we plan to extend the proposed framework to the continuous settings which is still an open challenge where only special cases are known [2], [17]. Another direction is multi-view learning via deep neural networks [56], [57], [58] where splitting methods can shed light on solver architectures with better parallelism and efficiency [59], [60].
Note that if $f \in C^1$ and $\nabla f(x)$ is $L$-Lipschitz continuous, then the function $f$ is said to be a $L$-smooth function.

Recall Definition 1, if a distribution is $\epsilon$-function then the smallest mass is strictly bounded away from zero by a positive constant $\epsilon$. The infimal measure is commonly assumed in non-parametric entropy/density estimation for the smoothness of the estimators [29], [30].

**Lemma 1:** Let $f(u) = \sum_{\mu_i} \mu_i \log \mu_i$ be the negative entropy function where two distinct measures $\mu, \nu$ are $\epsilon$-infinite. Then $f$ is $|\log \epsilon|\text{-Lipschitz continuous}$ and $1/\epsilon$-smooth.

**Proof:** The Lipschitz continuity follows as:

$$f(\mu) - f(\nu) = \sum_{x} \{\mu(x) - \nu(x)\} \log \frac{1}{\nu(x)} - D_{KL}(\mu \parallel \nu) \leq |\log \epsilon||\mu - \nu|.$$  

As for the smoothness:

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

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

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

With these results, we can establish similar smoothness conditions for the conditional entropy.

**Corollary 1:** Let $p_x$ be given, $p_{z|x}$ be $\epsilon$-infinite, then the conditional entropy $H(Z|X)$ is $|\log \epsilon|$-Lipschitz continuous and $1/\epsilon$-smooth.

**Proof:** Following Lemma 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 since:

$$H(Z^n|X) - H(Z^n|X) \leq |\log \epsilon\sum_{z} p(z)p(z^n|x) - p(z^n|x)|$$

$$\leq |\log \epsilon|\sup_{x \in X} p(x)||p_{z|x}^n - p_{z|x}^n||$$

$$= |\log \epsilon||p_{z|x}^n - p_{z|x}^n||.$$  

On the other hand, to prove the smoothness, similar to the r.h.s. of the inequality (32), we have:

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

**Definition 4:** A differentiable function $f : \mathbb{R}^n \rightarrow [0, +\infty)$ is said to be $\sigma$-hypoconvex, $\sigma \in \mathbb{R}$ if the following holds:

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

If $\sigma = 0$, (33) reduces to the definition of a convex function; $\sigma > 0$ corresponds to strong convexity whereas when $\sigma < 0$, it is known as the weak convexity [22], [27], [36].

A well-known example is the negative entropy function, which is 1-strongly convex in 1-norm [28], and consequently in 2-norm. Another example is the conditional entropy, which is weakly convex if the corresponding conditional probability mass is $\epsilon$-infinite as shown in the follow lemma.

**Lemma 2:** Let $G(p_{z|y}) = H(Z|Y)$. If $p_{z|y}$ is an $\epsilon_{z|x}$-infinite measure, then the function $G$ is $(2N_z N_y/\epsilon_{z|x})$-weakly convex, where $N_z = |Z|, N_y = |Y|$ denote the cardinalities of the random variables $Z, Y$, respectively.

**Proof:** See Appendix C.

A closely related concept to the hypoconvexity is the restricted weak convexity (Recall Definition 2). The restricted-weak convexity is adopted in our earlier work [16] to prove the convergence of an ADMM solver for IB. We further use this to prove the locally linear rate of convergence for Solver I. In the proposed Markovian framework, this concept relates to the data-processing inequality.

**Lemma 3:** Assume $p_{z|x}$ is $\epsilon_{z|x}$-infinite. Let $G(p_{z|x}) := -\gamma H(Z|X) + H(Z|Y)$ and $Y \rightarrow X \rightarrow Z$ forms a Markov chain. If $0 < \gamma < 1$, then for two $p_{z|x}^m, p_{z|x}^n \in \Omega_{z|x}$, where $\Omega_{z|x} := \{p(z|x)\} \sum_{x} p(z|x) = 1, \forall z \in Z, x \in X$, $G(p_{z|x})$ is $\gamma\omega_G$-restricted weakly convex.

$$G(p_{z|x}^m) - G(p_{z|x}^n) \geq \langle \nabla G(p_{z|x}^n), p_{z|x}^m - p_{z|x}^n \rangle$$

$$- \frac{\gamma\omega_G}{2} \|A_x p_{z|x}^m - A_x p_{z|x}^n\|^2,$$

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

**Proof:** see Appendix D.

Beyond smooth functions, if convexity applies as well, then we have the following descent lemma, commonly used in first-order optimization methods [26], [61], [62].

**Lemma 4 (Theorem 2.1.12 [62]):** If $f : \mathbb{R}^n \rightarrow [0, +\infty)$ is $\sigma$-strongly convex and $L$-smooth, then for any $x, y \in \mathbb{R}^n$, the following holds:

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

$$+ \frac{1}{\sigma + L}\|\nabla f(x) - \nabla f(y)\|^2.$$  

A recent result generalized the above to $\sigma$-hypoconvex $f$ which can be found in the reference therein [22]. Under $\epsilon$-infinity, the following results show that the positive entropy function is weakly convex.

**Lemma 5:** Given $p_x$, let $G(q)$ be defined as in (26). If $q = p_{z|x}$ is $\epsilon_{z|x}$-infinite, then $G(q)$ is $\sigma_G$-weakly convex w.r.t. $q$, where $\sigma_G := \max\{2|\beta - 1|N_z/\epsilon_{z|x}, 2N_z N_y/\epsilon_{z|x}\}$.

**Proof:** See Appendix B.

The following elementary identities are useful for the convergence proof. We list them for completeness.

$$2(u - w, v - w) = \|w - v\|^2 - \|u - v\|^2 - \|u - w\|^2.$$  

$$\|(1 - \alpha)u + \alpha v\|^2 = (1 - \alpha)\|u\|^2 + \alpha \|v\|^2 - \alpha(1 - \alpha)\|u - v\|^2.$$  

Lastly, by “linear” rate of convergence, we refer to the definition in [41].

**Definition 5:** 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

$$\frac{\|w^{k+1} - w^*\|}{\|w^k - w^*\|} \leq Q, \ \forall k > K_0.$$
On the other hand, the convergence of the sequence is \( R \)-linear if there is a \( Q \)-linearly convergent sequence \( \{\mu^k\}, \forall k \in \mathbb{N}, \mu^k \geq 0 \) such that:
\[
\|w^k - w^*\| \leq \mu^k, \forall k \in \mathbb{N}.
\]

B. Kurdyka-Łojasiewicz Inequality

To prove the convergence of alternating direction method of multiplier (ADMM) [18] for non-convex objectives, one can show that the corresponding augmented Lagrangian satisfies the Kurdyka-Łojasiewicz (KŁ) property [24], [25], [26]. Moreover, the rate of convergence can be determined in terms of the Łojasiewicz exponent of the augmented Lagrangian. This section is provided as a brief review of this tool.

The convergence for convex splitting methods is characterized through the Łojasiewicz inequality [24].

Definition 6: A function \( f(x) : \mathbb{R}^{|X|} \rightarrow \mathbb{R} \) is said to satisfy the Łojasiewicz inequality if there exists an exponent \( \theta \in [0, 1] \), \( \varepsilon > 0 \) and a critical point \( x^* \in \Omega^* := \{x|\nabla f(x) = 0, \forall x \in X\} \) with a constant \( C > 0 \), and a neighborhood \( \|x - x^*\| \leq \varepsilon \) such that:
\[
|f(x) - f(x^*)|^\theta \leq C \text{dist}(0, \nabla f(x)),
\]
where \( \text{dist}(a, \Omega) := \inf_{a \in \Omega} \|a - x\|_2 \).

Leveraging the main result in [24], the rate of convergence of splitting methods can be determined immediately by the associated Łojasiewicz exponents \( \theta \) if the objective function satisfies the KŁ property.

Definition 7: A function \( f(x) : \mathbb{R}^{|X|} \rightarrow \mathbb{R} \) is said to have the KŁ property if there exists a neighborhood around a stationary point \( x^* \) and a level set \( Q := \{x|\nabla f(x) < f(x^*) < f(x + \eta)\} \) with a margin \( \eta > 0 \) and a continuous concave function \( \varphi(s) : [0, \eta) \rightarrow \mathbb{R}_+ \), such that the following inequality holds:
\[
\varphi'(f(x) - f(x^*)) \text{dist}(0, \partial f(x)) \geq 1, \tag{37}
\]
where \( \partial f(x) \) denotes the sub-gradient [66] of \( f \) at \( x \) for non-smooth functions and reduces to gradient \( \nabla f \) for smooth functions.

Clearly, if \( \varphi(s) = Cs^{1-\theta} \), then (37) reduces to the Łojasiewicz inequality. The attribution to Kurdyka is due to the discovery of a variety of classical practices satisfying the KŁ property, the class of functions is said to have the \( o \)-minimal structure [43], i.e., sub-analytic and semi-algebraic functions. Once satisfying the KŁ property and knowing the exponent \( \theta \), the objective function, if solved with splitting methods, has the corresponding rate of convergence characterized depending on the value of \( \theta \). In particular, the most relevant case in the sequel, if the exponent \( \theta = 1/2 \) then the rate of convergence is locally linear around a neighborhood of a stationary point [24]. While the Łojasiewicz exponents of the \( o \)-minimal function class are often easy to calculate [25], [67], for more general functions, the exponents are difficult to determine [42], [68].

Recently, since the application of the KŁ inequality for convergence analysis of splitting methods in non-convex settings [25], [26], a wealth of optimization mathematics research has devoted to characterizing the convergence conditions under a assumed structure of the non-convex objective function. Among which, the one that is most relevant to ours is the (strongly) convex-weakly convex structure [21], [22], [27], [36]. The main discovery is that under this structure the convergence is assured if the penalty coefficient \( c \) is sufficiently large, characterized by the Lipschitz smoothness and properties of the operators in the linear constraints. We refer to [37] for a summary of convergence conditions of non-convex ADMM and [22] for Douglas-Rachford splitting (DRS) [23] and the references therein for recent advances.

C. 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, paired with Solver I:

Assumption A:

- There exists a stationary point \( w^* := (p^*, Bq^*, \nu^*) \) that belongs to a set \( \Omega^* := \{w|\nu \in \Omega, \nabla L_c = 0\} \).
- \( F(p) \) is \( L_{\rho} \)-smooth, \( \sigma_{\rho} \)-strongly convex while \( G(q) \) is \( L_{\sigma} \)-smooth and \( \omega_{\sigma} \)-restricted weakly convex.
- \( A \) is positive definite.
- The penalty coefficient \( c > c_{\min} \), and \( c_{\min} \) is defined as:
  \[
  c_{\min} := \max\{\omega_{G}, (L_{\rho} + \sigma_{\rho})b_A^2/\alpha\}.
  \]

The first-order optimization methods for (10), which gives the following minimizer conditions:
\[
\nu_{1/2}^{k+1} = \nu^k - (1 - \alpha)c(Ap^k - Bq^k),
\]
\[
\nabla F(p^{k+1}) = -A^T \nu_{1/2}^{k+1} - A^T (Ap^{k+1} - Bq^k) = -A^T \nu_{1/2}^{k+1},
\]
\[
\nu^{k+1} = \nu_{1/2}^{k+1} + c(Ap^{k+1} - Bq^k),
\]
\[
\nabla G(q^{k+1}) = B^T(\nu^{k+1} + c(Ap^{k+1} - Bq^k)). \tag{38}
\]

At a stationary point \( (p^*, q^*, \nu^*) \), the above reduces to:
\[
Ap^* = Bq^*, \quad \nabla F(p^*) = -A^T \nu^*, \quad \nu_{1/2}^{*} = \nu^*, \quad \nabla G(q^*) = B^T \nu^*. \tag{39}
\]

With the minimizer conditions, we present a sufficient decrease lemma for Solver I.

Lemma 6 (Sufficient Decrease I): Let \( \mathcal{L}_c \) be defined as in (9) and Assumption A is satisfied, then with Solver I,
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:
\[
\delta_p := \frac{\sigma_F L_p}{\mu_A^2(L_p + \sigma_F)} + c\left(1 - \frac{1}{\alpha}\right), \quad \delta_q := c - \omega_G, \quad \delta_\nu := \frac{1}{\mu_A^2(L_p + \sigma_F)} - \frac{1}{\alpha c},
\]
where \(\mu_A\) denotes the largest eigenvalue of the matrix \(A\).

**Proof:** See Appendix E.

By Lemma 6, the conditions that assure sufficient decrease are equivalent to the range of the penalty coefficient \(c\) and the relaxation parameter \(\alpha\) such that \(\delta_p, \delta_q, \delta_\nu\) are non-negative. When the conditions are satisfied, the sufficient decrease lemma implies the convergence of Solver I.

**Lemma 7 (Convergence I):** Suppose Assumption A is satisfied and \(0 < \alpha \leq 2\). Define the collective point at step \(k\) as \(w^k := (p^k, Bq^k, \nu^k)\), then the sequence \(\{w^k\}_{k \in \mathbb{N}}\) obtained from Solver I converges to a stationary point \(w^* \in \Omega^*\).

**Proof:** See Appendix F.

As a remark, the convergence is not point-wise. This can be observed as \(q\) in the collective point is pre-multiplied by the matrix \(B\). In practice, take IB for example, this corresponds to the symmetry of solutions [69], [70]. Nonetheless, point-wise convergence is not necessary as the mutual information, the metric involved in the information-theoretic optimization problems, is symmetric.

For Solver II, we develop a sufficient decrease lemma with more relaxed assumptions:

**Assumption B:**
- There exist stationary points \(w^* := (p^*, q^*, \nu^*)\) that belong to a set \(\Omega^* := \{w|w \in \Omega, \nabla \mathcal{L}_c = 0\}\).
- The function \(F(p)\) is \(L_p\)-smooth and convex while \(G(q)\) is \(L_q\)-smooth and \(\sigma_G\)-weakly convex.
- \(B\) is positive definite and \(A\) is full row rank.
- The penalty coefficient \(c\) satisfies:
  \[
c > \frac{\alpha \sigma_G + \sqrt{\alpha^2 \sigma_G^2 + 8(2 - \alpha)L_q^2 \mu_B^2}}{(4 - 2\alpha)\mu_B^2}.
\]

The corresponding first-order necessary conditions are:
\[
\nabla F(p^{k+1}) = -A^T[p^k + c(Ap^{k+1} - Bq^k)],
\]
\[
\nu^{k+1} = \nu^k - (1 - \alpha)c(Ap^{k+1} - Bq^k),
\]
\[
\nabla G(q^{k+1}) = B^T[q^{k+1} + c(Ap^{k+1} - Bq^{k+1})]
\]
\[
\nu^{k+1} = \nu^{k+1} + c(Ap^{k+1} - Bq^{k+1}).
\]

**Lemma 8 (Sufficient Decrease II):** Let \(\mathcal{L}_c\) be defined as in (9) and Assumption B is satisfied, then using Solver II, we have:
\[
\mathcal{L}_c(p^k, q^k, \nu^k) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu^{k+1}) \geq \frac{c}{2}\|Ap^k - Ap^{k+1}\|^2 - \frac{\sigma_G}{2}\|q^k - q^{k+1}\|^2 + c\left(1 - \frac{1}{\alpha}\right)\|Bq^k - Bq^{k+1}\|^2 - \frac{1}{\alpha c}\|\nu^k - \nu^{k+1}\|^2.
\]

**Proof:** See Appendix G.

In parallel to Lemma 7, we have the following convergence result for Solver II.

**Lemma 9 (Convergence II):** Suppose Assumption B 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 Solver II converges to a stationary point \(w^* \in \Omega^*\).

**Proof:** See Appendix H.

Note that the convergence of Solver II requires no strong convexity for the sub-objective function \(F(p)\). Moreover, the assumption for \(G(q)\) is more relaxed than that of Solver I, and hence the results apply to wider class of functions.

Another difference between the two assumption sets lies in the linear constraints. In the first assumption set, \(A\) is positive definite while \(B\) is positive definite in second set. In the Markovian information theoretic optimization problem (6), the linear constraints \(Ap - Bq\) represent the marginal/Markov relations of (conditional) probabilities. Therefore, only one of the two matrices \(A, B\) is identity, 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 the assumptions mentioned above hold. Inspired by [37], when \(F\) and \(G\) are further assumed to be Lipschitz continuous, we can relax the first assumption set but keep \(A\) to be positive definite as in the second set.

**Assumption C:**
- There exists a stationary point \(w^* := (p^*, q^*, \nu^*)\) that belongs to a set \(\Omega^* := \{w|w \in \Omega, \nabla \mathcal{L}_c = 0\}\).
- The function \(F(p)\) is \(L_p\)-smooth and convex while \(G(q)\) is \(L_q\)-smooth and \(\sigma_G\)-weakly convex.
- In addition, \(G(q)\) is \(M_q\)-Lipschitz continuous,
- \(A\) is positive definite and \(B\) is full row rank,
- The penalty coefficient \(c\) satisfies:
  \[
c > M_q \left[\chi_q + \sqrt{\chi_q^2 + 8(2 - \alpha)L_q^2 \lambda_B^2 \mu_{BB}^2}\right]/(4 - 2\alpha).
\]

where \(\chi_q := \sqrt{\chi_q^2 + 8(2 - \alpha)L_q^2 \lambda_B^2 \mu_{BB}^2}/(4 - 2\alpha)\).

When the above assumptions are imposed on Solver II, we have the following sufficient decrease lemma.

**Lemma 10 (Convergence III):** Suppose Assumption C is satisfied and \(0 < \alpha < 2\). Define \(w^k := (p^k, q^k, \nu^k)\) as the collective point at step \(k\), then the sequence \(\{w^k\}_{k \in \mathbb{N}}\) obtained from Solver II converges to a stationary point \(w^* \in \Omega^*\).

**Proof:** See Appendix I.

**D. Rate of Convergence Analysis**
In this part, we show that the rates of convergence of the algorithms, under the three sets of assumptions discussed in the previous part, are all locally linear. Specifically, the linear convergence is independent of the initialization. Moreover, the sequences obtained from the two algorithms, each converges to a local minimizer around its neighborhood [37], [63]. The results are based on the KL inequality that recently applied to the characterization of the rate of convergence for splitting
methods in non-convex problems [24], [25], [26], [61]. The analysis consists of two steps. First we show that (9), solved with either Solver I or Solver II, satisfies the KL property with a Łojasiewicz exponent \( \theta = 1/2 \). Then due to the following result, owing to [24], [26], and [42], we can prove the linear rate.

**Lemma 11 (Theorem 2 [24]):** Assume that a function \( \mathcal{L}_c(p, q, \nu) \) satisfies the KL property, define \( w_k \) the solution at step \( k \), and let \( \{w_k\}_{k \in \mathbb{N}} \) be a sequence generated by either Solver I or Solver II. Suppose \( \{w_k\}_{k \in \mathbb{N}} \) is bounded and the following relation holds:

\[
\left\| \nabla \mathcal{L}_c^k \right\| \leq C^* \left\| w_k - w_{k-1} \right\|,
\]

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 \( \theta \), then the following holds: (i)

1. If \( \theta = 0 \), the sequence \( \{w_k\}_{k \in \mathbb{N}} \) converges in a finite number of steps,
2. If \( \theta \in (0, 1/2] \) then there exist \( \tau > 0 \) and \( Q \in [0, 1) \) such that
   \[ \|w_k - w^\infty\| \leq \tau Q^k, \]
3. If \( \theta \in (1/2, 1) \) then there exists \( \tau > 0 \) such that
   \[ \|w_k - w^\infty\| \leq \tau k^{-\frac{1}{1-\theta}}. \]

**Proof:** See Appendix J, where we only include the case corresponding to \( \theta = 1/2 \) since it is the case that is relevant to the following discussion. For the proof for other scenarios, we refer the reader to [24].

The above result characterizes the rate of convergence in terms of the KL exponent, but except for certain types of functions, the calculation of the KL exponent is difficult. The following key result, due to [42], is useful in calculating the KL exponent of (9) and is included for completeness.

**Lemma 12 (Lemma 2.1 [42]):** Suppose that \( f \) is a proper closed function, \( \nabla f(\bar{w}) \neq 0 \). Then, for any \( \theta \in [0, 1] \), \( f \) satisfies the KL property at \( \bar{w} \) with an exponent of \( \theta \). In particular, define \( \eta := \frac{1}{2} \|\nabla f(\bar{w})\| > 0 \), then there exists \( \delta \in (0, 1) \) such that \( \|\nabla f(w)\| > \eta \) whenever \( \|w - \bar{w}\| < \varepsilon \) and \( f(\bar{w}) < f(w) < f(\bar{w}) + \delta \).

In literature, the KL inequality has been successfully adopted to find the rate of convergence for alternating algorithms such as ADMM and recently PRS or DRS [22]. In our case, it turns out that proving locally linear rate with the KL inequality only holds for \( 1 \leq \alpha \leq 2 \). As for \( 0 < \alpha < 1 \), inspired by the recent results that show locally R-linear rate of convergence for the primal ADMM [21], we adopt and extend the approach to Solver I and Solver II under the three sets of assumptions. Combining the two methods, we therefore show that the rates are locally linear for \( 0 < \alpha \leq 2 \).

**Lemma 13:** Let \( \mathcal{L}_c \) be defined as in (9) and let the sequence \( \{w_k\}_{k \in \mathbb{N}} \) obtained through either Solver I or Solver II be bounded. Denote \( \mathcal{L}^k_c := \mathcal{L}_c(p^k, q^k, \nu^k) \). 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 \( \mathcal{N} \) around a stationary point \( w^* \), such that \( \|w - w^*\| < \varepsilon \), \( \mathcal{L}_c^k \leq \mathcal{L}_c^* < \mathcal{L}_c^* + \delta \) with \( \delta, \varepsilon > 0 \), then \( \{\mathcal{L}_c^k\}_{k \in \mathbb{N}} \) is Q-linearly convergent and \( \{w_k\}_{k \in \mathbb{N}} \) converges R-linearly to \( w^* \) locally around \( \mathcal{N} \).

**Proof:** See Appendix K.

Remarkably, the rate of convergence with KL inequality is \( Q \)-linear, or in other words, monotonic convergence in terms of the error \( \|w_k - w_{k-1}\| \) in consecutive steps is guaranteed, while it is non-monotonic for \( R \)-linear rate. However, the weaker \( R \)-linear rate comes with milder assumptions imposed on the linear constraints, in particular, the full row rank assumptions are lifted.

In the rest of this part, we aim to prove that the sequence \( \{w_k\}_{k \in \mathbb{N}} \) obtained from any of the two proposed solvers satisfies the KL property. The results are based on the following lemmas. We start with a lemma developed for Solver I.

**Lemma 14:** Let \( \mathcal{L}_c \) be defined as in (9). For the sequence \( \{w_k\}_{k \in \mathbb{N}} \) obtained from Solver I where \( w_k := (p^k, Bq^k, \nu^k) \), if it is bounded and converges to a stationary point \( w^* \) satisfying (39), then we have:

\[
\mathcal{L}_c^{k+1} - \mathcal{L}_c^k \leq \chi_1 \|p^{k+1} - p^*\|^2 - \frac{c - \omega G}{2} \|Bq^{k+1} - Bq^*\|^2, \\
\|\nabla \mathcal{L}_c(w_{k+1})\| \geq (c^2 G + 1) \|Ap^{k+1} - Bq^{k+1}\|,
\]

where \( \chi_1 := \frac{c \chi_A}{2} - \sigma G \ell_p / (\ell_p + \sigma) \) and \( \mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k); \lambda_A, \mu_A \) denotes the largest and smallest eigenvalue of a matrix \( A \) respectively.

**Proof:** See Appendix L.

**Lemma 15:** Suppose Assumption A is satisfied, if the augmented Lagrangian (9) is solved with Solver I, then it satisfies the KL property with an exponent \( \theta = 1/2 \).

**Proof:** See Appendix M.

Given the exponent \( \theta = 1/2 \), then using Lemma 11, we prove that the rate of convergence is linear. For completeness, we include the statements of the main theorems below.

**Theorem 1:** 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 Solver I is bounded. Moreover, the sequence converges to a stationary point \( w^* \) at linear rate locally.

**Proof:** See Appendix N.

Similarly, for Solver II, we show that the Łojasiewicz exponent of the corresponding augmented Lagrangian is \( \theta = 1/2 \).

**Lemma 16:** Let \( \mathcal{L}_c \) be defined as in (9). For the sequence \( \{w_k\}_{k \in \mathbb{N}} \) obtained from Solver II, where \( w_k := (Ap^k, q^k, \nu^k) \), if the sequence is bounded and converges to a stationary point \( w^* \) satisfying (39), \( 0 < \alpha < 2 \), then we have:

\[
\mathcal{L}_c^{k+1} - \mathcal{L}_c^k \leq c |Ap^{k+1} - Bq^{k+1}|^2 + \frac{\sigma G}{2} \|q^{k+1} - q^*\|^2 \\
- \frac{c}{2} \|Bq^{k+1} - Bq^*\|^2 + \frac{c(2 - \alpha)}{2} \|Bq^k - Bq^*\|^2 \\
- \frac{c(2 - \alpha)}{2} \|Ap^{k+1} - Ap^*\|^2 + \frac{c(\alpha - 1)}{2} \|Ap^k - Ap^*\|^2,
\]

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

\[
\|\nabla \mathcal{L}_c(w_{k+1})\|^2 \geq \mu_W \left[ \|w_k - w_{k-1}\|^2 + c^2 \|Bq^k - Bq^{k+1}\|^2 - 2cLq \|q_k - q^{k+1}\|^2 \right],
\]

where \( \mu_W \) is the smallest eigenvalue of a matrix \( W \).
Proof: See Appendix O.

Lemma 17: Suppose Assumption B is satisfied and the matrix \( A \) is full row rank. For \( 0 < \alpha < 2 \), if the augmented Lagrangian (9) is solved with Solver II, then it satisfies the KL inequality with an exponent \( \theta = 1/2 \).

Proof: See Appendix P.

Observe that in Lemma 17, an additional full row rank assumption is imposed on the matrix \( A \). This is necessary to prove \( Q \)-linear rate of convergence with KL inequality. This method depends on the range of the relaxation parameter \( 1 \leq \alpha < 2 \). As for \( 0 < \alpha < 1 \), it turns out that we can show locally \( R \)-linear rate of convergence without assuming \( A \) to be full row rank.

Theorem 2: Suppose Assumption B is satisfied and the sequence \( \{w^k\} \) with \( w^k := (Ap^k, q^k, \nu^k) \) obtained from Solver II is bounded, then \( \{w^k\} \) converges to a stationary point \( w^* \) locally around a neighborhood \( ||w^* - w||^2 < \epsilon \) and \( L_\epsilon^* < L_\epsilon < L_\epsilon^* + \eta \) for some \( \epsilon, \eta > 0 \).

Proof: See Appendix Q.

Lastly, when imposing Assumption C on Solver II, we prove that the Łojasiewicz exponent in solving the augmented Lagrangian (9) is \( \theta = 1/2 \). Then we apply the KL inequality to prove that the rate of convergence is locally linear. The key difference of the Assumption C is that \( G \) is required to be Lipschitz continuous, which allows us to have inequalities such as \( ||q^m - q^n|| \leq M_q ||Bq^m - Bq^n|| \) with \( B \) not necessarily being positive definite [37]. We first adopt this result to prove \( \theta = 1/2 \).

Lemma 18: Suppose Assumption C is satisfied and the matrix \( B \) is full row rank. For \( 0 < \alpha < 2 \), if the augmented Lagrangian (9) is solved with Solver II, then it satisfies the KL property with an exponent \( \theta = 1/2 \).

Proof: See Appendix R.

Finally, using the above lemmas, we can prove locally linear rate of convergence for Solver II, paired with Assumption C.

Theorem 3: 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 \( \{w^k\} \) obtained from Solver II is bounded. Moreover, the sequence converges to a stationary point \( w^* \) at a linear rate locally.

Proof: See Appendix S.

APPENDIX B
PROOF OF LEMMA 5

Since in (26), \( G(p_{z|x}) = (\beta - 1)H(Z) + H(Z|X) \), 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 it is a scaled negative entropy function which is \((1-\beta)\)-strongly convex w.r.t. \( p_z \) and hence w.r.t. \( p_{z|x} \) as \( p_z = Q_x p_{z|x} \). 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 the strong convexity as a lower bound. On the other hand, if \( \beta > 1 \), for two distinct \( p_{z|x}^m, p_{z|x}^n \in \Omega_z \), we have:

\[
\langle \nabla H(Z^n), p_{z|x}^m - p_{z|x}^n \rangle - D_{KL}(p_{z|x}^m \parallel p_{z|x}^n) - \frac{1}{\varepsilon_z} || p_{z|x}^m - p_{z|x}^n ||_1^2
\]

\[
\geq \langle \nabla H(Z^n), p_{z|x}^m - p_{z|x}^n \rangle - \frac{N_z}{\varepsilon_{z|x}} || p_{z|x}^m - p_{z|x}^n ||_2^2,
\]

where the first inequality follows from the reversed Pinsker’s inequality due to the \( \varepsilon_z^{-1} \)-infimal assumption [71]. Then for the first term in the last inequality, due to the marginal relation \( Q_x p_{z|x} = p_z \), we have:

\[
\langle \nabla_x H(Z^n), p_{z|x}^m - p_{z|x}^n \rangle = \langle \nabla_x H(Z^n), p_{z|x}^m - p_{z|x}^n \rangle
\]

where \( \nabla_x \) 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 \( p_{z|x}^m - p_{z|x}^n \), we can simply discard the positive squared term.

APPENDIX C
PROOF OF LEMMA 2

For two arbitrary \( p_{z|x}^m, p_{z|x}^n \in \Omega_{z|x} \), consider the following:

\[
H(Z^m|Y) - H(Z^n|Y) = \sum_y p(y) \left[ \langle p_{z|x}^m - p_{z|x}^n, \log p_{z|x}^m \rangle - D_{KL}(p_{z|x}^m \parallel p_{z|x}^n) \right]
\]

\[
\geq \langle \nabla H(Z^m|Y), p_{z|x}^m - p_{z|x}^n \rangle - E_y \left[ \frac{1}{\varepsilon_{z|x}} || p_{z|x}^m - p_{z|x}^n ||_1^2 \right]
\]

\[
\geq \langle \nabla H(Z^n|Y), p_{z|x}^m - p_{z|x}^n \rangle - \frac{N_z}{\varepsilon_{z|x}} || p_{z|x}^m - p_{z|x}^n ||_2^2,
\]

where the first inequality follows the reverse Pinsker’s inequality [71] which holds when \( p_{z|x} \) is \( \varepsilon_{z|x}^{-1} \)-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 weak convexity, we complete the proof.

APPENDIX D
PROOF OF LEMMA 3

Since \( G(p_{z|x}) \) consists of two conditional entropy functions, the proof is divided into two parts. For the first part:

\[
- H(Z^m|X) + H(Z^n|X) = \sum_{x} p(x) \left\{ \sum_z |p(z^n|x) - p(z^m|x)| (\log p(z^n|x) + 1) \right\}
\]

\[
+ E_x[D_{KL}(p_{z|x}^m \parallel p_{z|x}^n)]
\]

\[
\geq \langle p_{z|x}^m - p_{z|x}^n, p(x) (\log p_{z|x}^m + 1) \rangle + D_{KL}(A_x p_{z|x}^m \parallel A_x p_{z|x}^n)
\]
where we use the log-sum inequality for the first and Pinsker's inequality for the second [28], followed by a 2-norm bound. As for the second part, ignore the trade-off parameter without loss of generality:

$$H(Z^n|Y) - H(Z^n) = \sum_y p(y) \left\{ \sum_z [p(z^n|y) - p(z^n)] \right\},$$

where we use the log-sum inequality for the first and Pinsker's inequality for the second [28], followed by a 2-norm bound.

Then for the second term in (43), through a similar technique in differential privacy [44], we have:

$$p(z^n|y) - p(z^n) \leq \frac{\zeta(y)}{p(y)} \sum_x p(x) [p(z^n|x) - p(z^n)],$$

where $\zeta(y) := \sup_{x \in X} p(y|x) - \inf_{x \in X} p(y|x)$.

Substituting the above into (43), we arrive at:

$$H(Z^n|Y) - H(Z^n) = \sum_{x} \left[ \sum_y \frac{\zeta^2(y)}{p(y)} \right] \|Ax_{z^n}|x\| - \|A_{z^n}p_{z^n}|x\|^2.$$

Combining the above with $\gamma$ pre-multiplied to the second part, it is clear that $G(z^n)$ satisfies the definition of restricted weak convexity, where $\omega := N_{z}N_{x} \zeta/\epsilon_{z|x} = \gamma$ and $\zeta := \sum_{y} \zeta^2(y)/p(y)$.

APPENDIX E

PROOF OF LEMMA 6

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

$$\mathcal{L}_c(p^{k}, q^{k}, \nu^{k}) - \mathcal{L}_c(p^{k}, q^{k+1}, 1/2) = -(\alpha - 1) c \|Ap^k - Bq^k\|^2.$$  (44)

Then for $p$-update, due to the $\sigma_F$-strong convexity and using Lemma 4, we have:

$$\mathcal{L}_c(p^{k}, q^{k}, \nu^{k+1}) - \mathcal{L}_c(p^{k+1}, q^{k}, \nu^{k+1}) \geq \langle \nabla G(p^{k+1}) + \nabla G(p^k) - \nabla G(p^{k+1}) \rangle + \frac{c}{2} \|Ap^{k+1} - Bq^k\|^2 + \frac{1}{L_p + \sigma_F} \|\nabla G(p^{k+1})\|^2 + \frac{\sigma_F L_p}{L_p + \sigma_F} \|p^{k+1} - p^k\|^2 + \frac{1}{L_p + \sigma_F} \|\nabla G(p^{k+1})\|^2$$

$$= \frac{c}{2} \|Ap^k - Ap^{k+1}\|^2 + \frac{\sigma_F L_p}{L_p + \sigma_F} \|p^k - p^{k+1}\|^2$$

where the first inequality is due to $\sigma_F$-strong convexity; the second is due to $A$ being positive definite. Then, for the dual update (fixing $p^{k+1}, q^k$), we have:

$$\mathcal{L}_c(p^{k+1}, q^{k}, \nu^{k+1}) - \mathcal{L}_c(p^{k+1}, q^{k+1}, \nu^{k}) \geq \langle \nabla G(p^{k+1}) + \nabla G(p^k) - \nabla G(p^{k+1}) \rangle + \frac{c}{2} \|Ap^{k+1} - Bq^k\|^2 + \frac{1}{L_p + \sigma_F} \|\nabla G(p^{k+1})\|^2 + \frac{\sigma_F L_p}{L_p + \sigma_F} \|p^{k+1} - p^k\|^2$$

where $\sigma_F := \mu_A^2/(L_p + \sigma_F) - 1/\alpha c$. Then by the positive definiteness of $A$, we have: $\|Ap^k - Ap^{k+1}\| \geq \mu_A \|p^k - p^{k+1}\|$, where $\mu_A$ denotes the smallest eigenvalue of $A$. Substitute this into (49), then we complete the proof.

APPENDIX F

PROOF OF LEMMA 7

By Assumption A, the coefficients $\delta_p, \delta_q, \delta_v$ defined in Lemma 7 are non-negative, so the next step is to show that $\{L_k\}_{k \in \mathbb{N}}$ is finite. Denote $L_k := \mathcal{L}_c(p^k, q^k, \nu^k)$ for simplicity. From Assumption A, we have:

$$\mathcal{L}_c - \mathcal{L}_c^N \geq \sum_{k=1}^{N} \|p^k - p^{k+1}\|^2 + \|\nu^k - \nu^{k+1}\|^2 + \|Bq^k - Bq^{k+1}\|^2,$$  (50)

where $C^* = \min \{|\delta_p, \delta_q, \delta_v| > 0\}$. Define the collective point at step $k$ as $w^k := (p^k, Bq^k, \nu^k)$, then since $\mathcal{L}_c$ is lower semi-continuous, for $N \rightarrow \infty$ and denote the limit point

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
w_\infty, \mathcal{L}_c^1 - \mathcal{L}_c^\infty \text{ is finite. This implies that the r.h.s. of (50) is finite. Then } \|\mathbf{w}_k - \mathbf{w}_{k+1}\|^2 \to 0 \text{ as } k \to \infty, \text{ since finite } \sum_1^\infty \|\mathbf{w}_k - \mathbf{w}_{k+1}\|^2 \text{ means boundedness of } \{\mathbf{w}_k\}_{k\in\mathbb{N}}. \text{ From this, we know that } w_\infty \in \Omega^*, \text{ or equivalently, for } k > N_0 \in \mathbb{N} \text{ sufficiently large, } w_k \to w^* \text{ as } k \to \infty, \text{ which proves that } \{\mathbf{w}_k\}_{k\in\mathbb{N}} \text{ converges to } w^*.

**APPENDIX G**

**PROOF OF LEMMA 8**

First, by the convexity of $F$:

\[
\mathcal{L}_c(p_k, q_k, \nu_k) - \mathcal{L}_c(p_{k+1}, q_k, \nu_k) \geq \langle \nabla F(p_{k+1}) + A^T \nu_k, p_k - p_{k+1} \rangle + \frac{c}{2} \|A p_k - B q_k\|^2 - \frac{c}{2} \|A p_{k+1} - B q_k\|^2 \\
= -c \langle A p_{k+1} - B q_k, A p_k - A p_{k+1} \rangle + \frac{c}{2} \|A p_k - B q_k\|^2 - \frac{c}{2} \|A p_{k+1} - B q_k\|^2 \\
= \frac{c}{2} \|A p_k - A p_{k+1}\|^2, \tag{51}
\]

where the last equality is due to the minimizer conditions (40). Then for the relaxation step (11b), where $p_{k+1}, q_k$ are fixed:

\[
\mathcal{L}_c(\nu_k) - \mathcal{L}_c(\nu_{k+1}/2) = -(\alpha - 1)c \|A p_{k+1} - B q_k\|^2. \tag{52}
\]

On the other hand, by the $\sigma_G$-weak convexity of $G$, we have the following lower bound:

\[
\mathcal{L}_c(p_{k+1}, q_k, \nu_{k+1}/2) - \mathcal{L}_c(p_{k+1}, q_{k+1}, \nu_{k+1}/2) \geq \langle \nabla G(q_{k+1}) - B^T \nu_{k+1}/2, q_k - q_{k+1} \rangle - \frac{\sigma_G}{2} \|q_k - q_{k+1}\|^2 \\
+ \frac{c}{2} \|A p_{k+1} - B q_k\|^2 - \frac{c}{2} \|A p_{k+1} - B q_{k+1}\|^2 \\
= c \langle A p_{k+1} - B q_k, B q_k - B q_{k+1} \rangle - \frac{\sigma_G}{2} \|q_k - q_{k+1}\|^2 \\
+ \frac{c}{2} \|A p_{k+1} - B q_k\|^2 - \frac{c}{2} \|A p_{k+1} - B q_{k+1}\|^2 \\
= \frac{c}{2} \|B q_k - B q_{k+1}\|^2 - \frac{\sigma_G}{2} \|q_k - q_{k+1}\|^2. \tag{53}
\]

Lastly, for the dual ascend (11d), where $p_{k+1}, q_{k+1}$ are fixed:

\[
\mathcal{L}_c(\nu_{k+1}/2) - \mathcal{L}_c(\nu_{k+1}) = -c \|A p_{k+1} - B q_k\|^2. \tag{54}
\]

Combining (52) and (54) using the identity (36), we get:

\[
\frac{1}{c \alpha} \|\nu_k - \nu_{k+1}\|^2 = c \|A p_{k+1} - B q_{k+1}\|^2 + c(\alpha - 1) \|A p_{k+1} - B q_k\|^2 - c(1 - \frac{1}{\alpha}) \|B q_k - B q_{k+1}\|^2. \tag{55}
\]

Summing (51), (53) and (55), we get:

\[
\mathcal{L}_c(p_k, q_k, \nu_k) - \mathcal{L}_c(p_{k+1}, q_{k+1}, \nu_{k+1}) \geq \frac{c}{2} \|A p_k - A p_{k+1}\|^2 + c \left( \frac{1}{\alpha} - \frac{1}{2} \right) \|B q_k - B q_{k+1}\|^2 - \frac{\sigma_G}{2} \|q_k - q_{k+1}\|^2 - \frac{1}{\alpha c} \|\nu_k - \nu_{k+1}\|^2, \tag{56}
\]

which completes the proof.

**APPENDIX H**

**PROOF OF LEMMA 9**

By assumption, $B$ is positive definite. Denote its smallest eigenvalue as $\mu_B$ and $\mathcal{L}_c^k := \mathcal{L}_c(p_k, q_k, \nu_k)$ for simplicity, then 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 $\|B^{-1}\| = \mu_B$. On the other hand, for the dual variable, we have:

\[
\|\nu_k - \nu_{k+1}\| = \|(B^{-T}B^T)(\nu_k - \nu_{k+1})\| \leq \|B^{-T}\| \|\nabla G(q_k) - \nabla G(q_{k+1})\| \leq L_q \|B^{-T}\| \|q_k - q_{k+1}\|,
\]

combining the two results above, we have the following lower bound to Lemma 8:

\[
\mathcal{L}_c^k - \mathcal{L}_c^{k+1} \geq \left[ c \mu_B^2 \left( \frac{1}{\alpha} - \frac{1}{2} \right) - \frac{\sigma_G}{2} \frac{L_q \mu_B^2}{\alpha c} \right] \|q_k - q_{k+1}\|^2 + \frac{c}{2} \|A p_k - A p_{k+1}\|^2. \tag{57}
\]

Then, we would like to make the coefficient of the squared norm $\|q_k - q_{k+1}\|^2$ be positive. Through quadratic programming, we have the desired range of the penalty coefficient $c$:

\[
c > \frac{\alpha \sigma_G + \sqrt{\alpha^2 \sigma_G^2 + 8(2 - \alpha) L_q^2 \mu_B^4}}{4 - 2 \alpha}, \tag{58}
\]

which is satisfied as listed in Assumption B. Then, for any $c$ satisfying (57), define $c^* := c/2$, we have:

\[
\mathcal{L}_c^k - \mathcal{L}_c^{k+1} \geq c^* \left[ \|q_k - q_{k+1}\|^2 + \|A p_k - A p_{k+1}\|^2 \right].
\]

Then, consider the following:

\[
\mathcal{L}_c^k - \mathcal{L}_c^N \geq c^* \sum_{i=1}^{N-1} \|q_k - q_{k+1}\|^2 + \|A p_k - A p_{k+1}\|^2. \tag{59}
\]

Define the collective point at step $k$ as $w_k := (A p_k, q_k, \nu_k)$. Since $\mathcal{L}_c$ is lower semi-continuous, the l.h.s. of (58) is finite. Note that the r.h.s. of (58) does not depend on the dual variable $\nu$, we can further define a condensed point at step $k$ as $z_k := (A p_k, q_k)$. Then, by letting $N \to \infty$ and denoting the limit point $z_\infty$, since $\mathcal{L}_c^1 - \mathcal{L}_c^\infty$ is finite, the r.h.s. of (58) is finite. This implies $\|z_k - z_{k+1}\|^2 \to 0$ as $k \to \infty$, since $\sum_{k \in \mathbb{N}} \|z_k - z_{k+1}\|^2$ is a Cauchy sequence. From this we know that $z_\infty = z^*$. Moreover, due to (20), $L_q^2 \mu_B^2 \|q_k - q_{k+1}\|^2 \geq \|\nu_k - \nu_{k+1}\|^2$ and hence $\|\nu_k - \nu_{k+1}\|^2 \to 0$ as $k \to \infty$. Therefore $\nu_\infty = \nu^*$. So, together we have $w_k \to w^*$ as $k \to \infty$ which proves that $\{w_k\}_{k \in \mathbb{N}}$ converges to $w^*$.

**APPENDIX I**

**PROOF OF LEMMA 10**

Following the steps (51), (53) and (55) in Appendix G, we start from (56). Define $\mathcal{L}_c^k := \mathcal{L}_c(p_k, q_k, \nu_k)$ as the function value evaluated with step $k$ solutions:

\[
\mathcal{L}_c^k - \mathcal{L}_c^{k+1}
\]
Therefore

\[ L = \frac{1}{\alpha - 1} \left( Bq^k - Bq^{k+1} \right)^2 - \frac{1}{\alpha c} \mu_k - \nu^{k+1} \right)^2 \]

\[ \geq c \mu_k^2 \frac{1}{\alpha - 1} \left( Bq^k - Bq^{k+1} \right)^2 - \frac{1}{\alpha c} \mu_k - \nu^{k+1} \right)^2 \]

where \( \kappa_M \) is defined as:

\[ \kappa_M := \left[ \frac{c}{Mq} \left( \frac{1}{\alpha - 1} - \sigma_G \frac{L^2 \lambda_B^4 B^2}{\alpha c} \right) \right]. \]

In the last inequality, the first term is due to \( A \) being positive definite, and for the second term, we follow [37] and use Lipschitz continuity of \( G \) to have \( \| q^k - q^{k+1} \| \leq M \| Bq^k - Bq^{k+1} \| \); we denote \( \lambda_B := \| B \| \) as the largest singular value of a matrix \( B \) and \( \mu_B \) for the smallest eigenvalue of \( B \); For \( \| \nu^k - \nu^{k+1} \| \), since \( B \) is full row rank and \( G \) is \( L_0 \)-smooth, we have:

\[ \| \nu^k - \nu^{k+1} \| = \left\| \left( BB^T \right)^{-1} BB^T \left( \nu^k - \nu^{k+1} \right) \right\| \]

\[ \leq \mu_B \lambda_B^2 \left\| \nabla G(q^k) - \nabla G(q^{k+1}) \right\| \]

\[ \leq \mu_B \lambda_B^2 L_0 \left\| \nu^k - \nu^{k+1} \right\|. \]  (60)

From elementary quadratic programming, the range in terms of the penalty coefficient \( c \) that assures that the second term of the last inequality in (59) is positive:

\[ c > Mq \left[ \frac{Mq \sigma_G \alpha + \left( Mq \sigma_G \alpha \right)^2 + 8(2 - \alpha) L^2 \lambda_B^4 \mu_B^2 B^2}{4 - 2 \alpha} \right]. \]

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:

\[ L_{c-k}^k \geq \tau^* \left( \| p^k - p^{k+1} \|^2 + \| q^k - q^{k+1} \|^2 \right) \]

Then denote \( w^k := (p^k, q^k, \nu^k) \) the collective point at step \( k \); \( L_{c-k}^k \) is the function value evaluated with \( w^k \). Summing both sides of the inequality (60), we have:

\[ L_{c-k}^k - L_{c-k}^n \geq \tau^* \sum_{k=1}^{N-1} \left( \| p^k - p^{k+1} \|^2 + \| q^k - q^{k+1} \|^2 \right) \]

By assumption the l.h.s. of the above inequality is lower semi-continuous and therefore is finite. So as \( N \to \infty \), \( L_{c-k}^1 - L_{c-k}^\infty \leq +\infty \). This implies that the r.h.s. is finite and therefore \( \| p^k - p^{k+1} \|^2 \rightarrow 0 \) and \( \| q^k - q^{k+1} \|^2 \rightarrow 0 \) as \( k \to \infty \). Due to (60), we know that \( \| p^k - p^{k+1} \|^2 \rightarrow 0 \) 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 \( \{ w^k \}_{k \in \mathbb{N}} \) converges to \( w^* \).

APPENDIX J

PROOF OF LEMMA 11

Denote \( L_{c-k}^k := L_{c-k}(p^k, q^k, \nu^k) \), without loss of generality let \( L_{c-k}^* = 0 \), 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 \( \Phi \) (Note that the gradient is evaluated at \( w^k \)):

\[ (L_{c-k}^k)^{1-\theta} - (L_{c-k}^{k+1})^{1-\theta} \geq (1 - \theta) \left( L_{c-k}^k \right)^{1-\theta} [L_{c-k}^k - L_{c-k}^{k+1}] \]

\[ \geq C_0^\theta \left( L_{c-k}^k \right)^{1-\theta} \| w^k - w^{k+1} \|^2 \]

\[ \geq C_0^\theta \| \nabla L_{c-k}^k \|^{-1} \| w^k - w^{k+1} \|^2, \]  (61)

for some constant \( C_0^\theta > 0 \). The second inequality is due to Lemma 7 and the last inequality is due to Lemma 15.

Then, by assumption, for some constant \( C^* > 0 \), we have:

\[ \| \nabla L_{c-k}^k \| \leq C^* \| w^k - w^{k+1} \|. \]  (62)

Substitute the above into (61), define \( C_1 := C^*/C^*(1 - \theta) \), we get:

\[ (L_{c-k}^k)^{1-\theta} - (L_{c-k}^{k+1})^{1-\theta} \geq C_1 \| w^k - w^{k+1} \|^2 \]

From the above, we have:

\[ \| w^{k+1} - w^k \| \leq \| w^k - w^{k+1} \| + C_2 \left[ (L_{c-k}^k)^{1-\theta} - (L_{c-k}^{k+1})^{1-\theta} \right] \]

\[ \| w^k - w^{k-1} \| + C_3 \| \nabla L_{c-k}^k \|^{-\theta} \]

\[ \leq \| w^{k+1} - w^k \| + C_4 \| w^k - w^{k-1} \|^{-\theta}, \]  (63)

where we define \( C_2 := \sqrt{1/(2C_1)} \), \( C_3 := C_2(C_0)^{1/\theta}, C_4 := C_2(C^*)^{1/\theta} \). For the first inequality, we use the identity \( 2a^2 + b^2 \); the second inequality is due to the non-increasing sequence \( \{ L_{c-k}^k \}_{k \in \mathbb{N}} \); the third inequality is due to the KL property, and the last inequality follows (62). Then, by defining \( \Delta_k := \sum_{i=k}^{\infty} \| w^{i+1} - w^i \| \), and summing both sides of (63) with \( k \in \mathbb{N} \), we have:

\[ \Delta_k \leq (\Delta_{k-1} - \Delta_k) + (\Delta_{k-1} - \Delta_k)^{1-\theta}. \]  (64)

Finally, from Lemma 15, \( \theta = 1/2 \), 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. That is, the Cauchy sequence \( \Delta_k \) converges \( Q \)-linearly fast.

APPENDIX K

PROOF OF LEMMA 13

By assumption, denote \( \Delta_{c-k}^k := L_{c-k}^k - L_{c-k}^* \), we have:

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

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

\[ \frac{\Delta_{c-k}^{k+1}}{\Delta_{c-k}^k} \leq \frac{K^*}{1 + K^*} \frac{K^* c^2}{1 + K^*} \frac{1}{\Delta_{c-k}^{k+1}} \]

\[ \leq \frac{K^*}{1 + K^*} \frac{K^* c^2}{1 + K^*} \frac{1}{\Delta_{c-k}^{k+1}} \]

where the second inequality follows from the sufficient descent lemma and by definition, \( \delta > \Delta_{c-k}^{k+1} - L_{c-k}^* > \xi > 0 \), as \( w^{k+1} \notin \Omega^* \). Therefore, we can simply choose \( \epsilon < \sqrt{\xi / K^*} < \sqrt{\delta / K^*} \), which shows that the convergence of the sequence values \( \{ L_{c-k}^k \}_{k \in \mathbb{N}} \) is \( Q \)-linear locally around the
neighborhood of a stationary point \( w^* \). In turns, we have for \( n > N_0 \in \mathbb{N} \):
\[
\rho_p \| A p^n - A p^{n+1} \|^2 \leq L^\gamma - L^{\gamma+1} \leq K_p Q^n,
\]
\[
\rho_q \| q^n - q^{n+1} \|^2 \leq L^\nu - L^{\nu+1} \leq K_q Q^n,
\]
\[
\rho_b \| \nu^n - \nu^{n+1} \|^2 \leq L_b^\nu - L_b^{\nu+1} \leq K_b Q^n,
\]
for some \( K_p, K_q, K_\nu \) are positive and \( 0 < Q < 1 \). Combine the above together, we have:
\[
\tilde{\rho} \| w^n - w^{n+1} \|^2 \leq \tilde{K} Q^n,
\]
where \( \tilde{K} = K_p + K_q + K_\nu \) and \( \tilde{\rho} = \min \{ \rho_p, \rho_q, \rho_b \} \). Now, we have the sequence \( \{ w^n \} \) as \( n \to \infty \), we get:
\[
\| w^n - w^* \|^2 \leq \frac{\tilde{K} Q^n}{\tilde{\rho} (1 - Q)},
\]
and hence \( \{ w^n \} \) is R-linearly convergent.

**APPENDIX L
PROOF OF LEMMA 14**

By the definition in (9) along with the properties of \( F \) and \( G \), using Solver I with the first order necessary conditions (38), and denote \( L^k_c := L_c(p^k, q^k, \nu^k) \) for simplicity, we have:
\[
L^{k+1}_c - L_c \leq \langle \nabla F(p_k^{k+1}, p^{k+1} - p) - \frac{\sigma F L_p}{L_p + \sigma F} \| p^{k+1} - p \|^2 
+ \langle \nabla G(q_k^{k+1}, q^{k+1} - q); \nu, A p - B q 
+ \frac{c}{2} \| A p^{k+1} - B q^{k+1} \|^2 \leq L^\gamma - L^{\gamma+1} \leq K_p Q^n,
\]
where the last equality follows from (38). By showing that \( \| L^{k+1}_c \|^2 \geq K \| A p^{k+1} - B q^{k+1} \|^2 \) with \( K := c^2 \mu_A^2 + 1 \) where \( \mu_A \) is the smallest eigenvalue of the matrix \( A \), we complete the proof.

**APPENDIX M
PROOF OF LEMMA 15**

By assumption, \( c > \omega c \) and by Cauchy-Schwarz inequality
\[
\| u - v \|^2 \geq (1 - 1/t) \| u \|^2 + (1 - t) \| v \|^2,
\]
then consider
\[
\frac{c - \omega c}{K} = \frac{t - 1}{1 - t} \Rightarrow K = \frac{c - \omega c}{t},
\]
for some \( t > 1 \), the following holds:
\[
L^{k+1}_c - L_c^* \leq c_G \| p^{k+1} - p^* \|^2 + \| A p^{k+1} - B q^{k+1} \|^2,
\]
where we define the following:
\[
c_G := \frac{1}{2} \left( \frac{c}{2} \left( \frac{\sigma F L_p}{L_p + \sigma F} \| p^{k+1} - p \|^2 \right) + \frac{c}{2} \| A p^{k+1} - B q^{k+1} \|^2 \right),
\]
and hence \( c_G \geq 0 \) for some c that is sufficiently large. Substitute the above into Lemma 14, we have:
\[
L^{k+1}_c - L_c^* \leq c_G c^2 \frac{1}{K_1} \| \nabla L^{k+1}_c \|^2
\leq \frac{c_G c^2}{K_1} \left( \frac{1}{4} \| \nabla L^k_c \|^2 + \frac{1}{K_1} \right),
\]
where \( K_1 := c^2 \mu_A^2 + 1 > 0 \); the first inequality is due to Lemma 14 and \( \| w^{k+1} - w^* \| < \epsilon \) around the neighborhood of \( w^* \); the last inequality follows from Lemma 12. By taking square root of both sides we complete the proof.

**APPENDIX N
PROOF OF THEOREM 1**

The convergence follows the sufficient decrease lemma (Lemma 7), so proving the rate of convergence suffices. Due to Lemma 15, \( L_c(p, q, \nu) \) satisfies the KL property with an exponent \( \theta = 1/2 \). Then for the gradient norm \( \| \nabla L_c \| \), by Lemma 14 we have:
\[
\| \nabla L^k_c \| \leq c_a \| A p^k - B q^k \|
\leq c_a \left( \| A p^k - B q^{k-1} \| + \| B q^k - B q^{k-1} \| \right),
\]
where \( c_a := c + \lambda_A \). Then, suppose \( 1 \leq \alpha \leq 2 \), by (47):
\[
\| A p^k - B q^{k-1} \|^2 \leq \frac{1}{\alpha^2} \| p^k - \nu^{k-1} \|^2 + (1 - \frac{1}{\alpha}) \| A p^k - A p^{k-1} \|^2.
\]
Substitute the above into (66), we get:
\[
\| \nabla L^k_c \| \leq c^*_a \left( \| p^k - \nu^{k-1} \|^2 + \| p^k - p^{k-1} \|^2 + (1 - \frac{1}{\alpha}) \| A p^k - A p^{k-1} \|^2 \right),
\]
where \( c^*_a := \max \{ c_a / (\alpha^2 c^2), c_a \lambda_A \} \). This shows that for \( 1 \leq \alpha \leq 2 \), and some constant \( c_{\alpha t} > 0 \), we have:
\[
\| \nabla L^k_c \| \leq c_{\alpha t} \| w^k - w^{k-1} \|.
\]
Then by Lemma 11 along with $\theta = 1/2$, we prove the locally linear rate of convergence for the case $1 \leq \alpha \leq 2$. As for $0 < \alpha < 1$, from Lemma 7 and Assumption A, there exists a constant $K' > 0$ such that:

$$
L_{c}^{k} - L_{c}^{k+1} \geq K^{*} \|w^{k} - w^{k+1}\|^2.
$$

Moreover, denote $w^{*} := (p^{*}, Bq^{*}, \nu^{*})$ as a stationary point, due to Lemma 14, we have:

$$
L_{c}^{k+1} - L_{c}^{*} \leq \left(\frac{c\lambda^{2}}{2} - \frac{\sigma_{p} L_{p}}{L_{p} + \sigma_{F}}\right) \|p^{k+1} - p^{*}\|^2
$$

$$
- \frac{c - \omega_{G}}{2} \|Bq^{k+1} - Bq^{k}\|^2.
$$

By Assumption A, there always exists constants $K_{1}, K_{1}^{'}, K > 0$ and a neighborhood around the stationary point $w^{*}$ such that:

$$
L_{c}^{k+1} - L_{c}^{*} \leq K_{1} \|w^{k+1} - \nu^{*}\|^2
$$

$$
\leq K_{1}' \|w^{k} - w^{*}\|^2 + K_{1}' \|w^{k+1} - w^{k}\|^2
$$

$$
\leq K_{1}' \|w^{k+1} - \nu^{*}\|^2 + \frac{K_{1}'}{K} (L_{c}^{k} - L_{c}^{k+1}).
$$

Then using Lemma 13, we conclude that the sequence \{w^{k}\}_{k>0}, N_{0} \in \mathbb{N} converges R-linearly to $w^{*}$. We conclude that for the full range of $0 < \alpha \leq 2$, the rate of convergence is locally linear.

**APPENDIX O

**PROOF OF LEMMA 16**

For the first part, denote $L_{c}^{k} := L_{c}(p^{k}, q^{k}, \nu^{k})$ the function value evaluated with step $k$ solutions, we have:

$$
L_{c}^{k+1} - L_{c}^{*} \leq \langle \nabla F(p^{k+1}), p^{k+1} - p \rangle + \langle \nabla G(q^{k+1}), q^{k+1} - q \rangle
$$

$$
+ \langle \nu^{k+1}, Ap^{k+1} - Bq^{k+1} \rangle + \frac{\sigma_{G}}{2} \|q^{k+1} - \nu^{k}\|^2
$$

$$
- \langle \nu^{k} - \nu^{k+1}, Ap^{k} - Bq^{k+1} \rangle + \frac{c}{2} \|Ap^{k+1} - Bq^{k+1}\|^2
$$

$$
- \frac{c}{2} \|Ap^{k+1} - Bq^{k+1}\|^2
$$

$$
- \frac{c}{2} \|Ap^{k+1} - Bq^{k+1}\|^2 + \frac{\sigma_{G}}{2} \|q^{k+1} - q^{k}\|^2
$$

$$
+ c(\nu^{k+1} - \nu^{k}, Ap^{k+1} - Bq^{k+1}) - c(2 - \alpha)(Ap^{k+1} - Bq^{k+1}, Ap^{k+1} - Ap),
$$

(68)

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

$$
c(\nu^{k+1} - \nu^{k+1}, Ap^{k+1} - Bq^{k+1})
$$

$$
- c(2 - \alpha)(Ap^{k+1} - Bq^{k+1}, Ap^{k+1} - Ap)
$$

$$
= \frac{c}{2} \|Ap^{k+1} - Bq^{k+1}\|^{2} + \|Ap^{k+1} - Bq^{k+1}\|^{2}
$$

$$
+ \|Ap^{k+1} - Ap\|^{2}
$$

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

$$
+ \|Ap^{k+1} - Ap\|^{2}.
$$

(69)

Substitute (69) into (68), using identities (35) and (36), and let $w = w^{*}$, so that $Ap^{*} = Bq^{*}$, we have:

$$
L_{c}^{k+1} - L_{c}^{*}
$$

$$
\leq c \|Ap^{k+1} - Bq^{k+1}\|^{2} - \frac{c}{2} \|Bq^{k+1} - Bq^{*}\|^{2}
$$

$$
+ \frac{\sigma_{G}}{2} \|q^{k+1} - q^{*}\|^{2} + c(\alpha - 1) \|Ap^{k+1} - Ap^{*}\|^{2}
$$

$$
+ \frac{c(2 - \alpha)}{2} \|Bq^{k} - Bq^{*}\|^{2}
$$

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

(70)

By identity (36) and the minimizer conditions (40):

$$
c \|Ap^{k+1} - Bq^{k+1}\|^{2}
$$

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

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

(71)

Substitute the above into (70), we have:

$$
L_{c}^{k+1} - L_{c}^{*}
$$

$$
\leq \frac{1}{\alpha} \|\nu^{k} - \nu^{k+1}\|^{2} + \frac{\sigma_{G}}{2} \|q^{k+1} - q^{k}\|^{2}
$$

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

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

$$
+ c \left(1 - \frac{1}{\alpha}\right) \|Bq^{k} - Bq^{k+1}\|^{2},
$$

(72)

which completes the proof for the first part. As for the second part, consider:

$$
\nabla L_{c}^{k+1} = \nabla F(p^{k+1}) + A^{T} [\nu^{k+1} + c (Ap^{k+1} - Bq^{k+1})]
$$

$$
- \nabla G(q^{k+1}) - B^{T} [\nu^{k+1} + c (Ap^{k+1} - Bq^{k+1})] - c B^{T} (Ap^{k+1} - Bq^{k+1}) - c B^{T} (Ap^{k+1} - Bq^{k+1})
$$

(73)

Denote the smallest positive eigenvalue of a matrix $W$ as $\mu_{W}$, by assumption, since $AA^{T} > 0$, we have:

$$
\|\nabla L_{c}^{k+1}\|^{2}
$$

$$
\geq \mu_{AA^{T}} (\|\nu^{k} - \nu^{k+1}\|^{2} + c^{2} \|Bq^{k} - Bq^{k+1}\|^{2}
$$

$$
- 2c(\nu^{k} - Bq^{k+1}, Bq^{k} - Bq^{k+1})
$$

$$
+ (2c B^{T} + 1) \|Ap^{k+1} - Bq^{k+1}\|^{2}
$$

$$
\geq \mu_{AA^{T}} (\|\nu^{k} - \nu^{k+1}\|^{2} + c^{2} \|Bq^{k} - Bq^{k+1}\|^{2}
$$

$$
- 2c L_{q} \|q^{k} - q^{k+1}\|^{2},
$$

(73)

where in the last inequality, we use the minimizer condition (40) and $L_{q}$-smoothness of $G$. Authori
From Lemma 16, discarding the terms with negative coefficients, using the Cauchy-Schwarz inequality on \(|Bq^k - Bq^*|^2|\) and denote \(\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)\), we have:

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

where the inequality follows \(||u + v||^2 \leq 2(||u||^2 + ||v||^2)\), applied to \(||Bq^k - Bq^*||^2\). Note that the coefficient of the term \(||Bq^k - Bq^{k+1}||^2\) is due to \(\alpha + 1/\alpha \geq 2\). Then by defining \(2C_G := \left[\sigma_G^2 - c\lambda_B^2(1/\mu_B^2 + 4 - 2\alpha)\right]^{1/2}\), where \([1]^+ := \max\{0,1\}\), we have:

\[
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \frac{1}{\alpha c} ||\nu^k - \nu^{k+1}||^2 + c\lambda_B^2 ||q^k - q^{k+1}||^2 \\
+ \frac{c}{2} ||Ap^{k+1} - Ap^*||^2 + C_G||q^{k+1} - q^*||.  \tag{75}
\]

On the other hand, since \(B\) is positive definite by assumption, we can further find a lower bound of (73):

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

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

\[
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \leq \frac{K_2}{K_1} ||\nabla \mathcal{L}_c^{k+1}||^2 + K_3 \left(||Ap^{k+1} - Ap^*||^2 \\
+ ||q^{k+1} - q^*||^2\right) \\
\leq \frac{K_2}{K_1} ||\nabla \mathcal{L}_c^{k+1}||^2 \left(1 + \frac{3\epsilon_c^2}{2}\right),
\]

where \(K_3 := \max\{c\alpha - 1/2, C_G\}\) and the last inequality follows Lemma 12, that is, around a neighborhood of \(w^*\) with \(||w - w^*|| < \epsilon\) there exists \(\eta > 0\) 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 \(\theta = 1/2\).

### APPENDIX Q

**Proof of Theorem 2**

Denote \(\mathcal{L}_c^k := \mathcal{L}_c(p^k, q^k, \nu^k)\) for simplicity. From Lemma 8, there always exists a stationary point \(w^* := \langle Ap^*, q^*, \nu^*\rangle\) that the sequence \(\{\mathcal{L}_c^k\}\) is converging to. By assumption, the penalty coefficient \(c\) is large enough such that the sufficient decrease lemma holds, which implies convergence. As for the corresponding rate, for \(1 \leq \alpha < 2\), by Lemma 17, the KL exponent \(\theta = 1/2\). In addition, from (72) we have:

\[
\|\nabla \mathcal{L}_c^{k+1}\|^2 \\
\leq \|A^T[\nu^{k+1} - \nu^* + c(Bq^k - Bq^*)]\|^2 \\
+ (1 + c^2\lambda_B^2) \|Ap^{k+1} - Ap^*\|^2 \\
\leq 2\lambda_{AA^T} (||\nu^{k+1} - \nu^*||^2 + \alpha Bq^{k+1} - Bq^k)_2^2 \\
+ (1 + c\lambda_B^2) \|Ap^{k+1} - Ap^*\|^2,
\]

where the inequality follows as \(A\) is full row rank. Then using the identity (36), and the minimizer conditions (40), we have:

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

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

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

For the second term, since \(B\) is positive definite, \(\|Bq^{k+1} - Bq^k\|^2 \leq \lambda_B^2 \|q^{k+1} - q^k\|^2\). Substitute this 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:

\[
\|\nabla \mathcal{L}_c^{k+1}\|^2 \leq M^* ||w^{k+1} - w^k||^2. \tag{76}
\]

Substitute (62) with (76), and by Lemma 11, the sequence \(\{w^k\}_{k \in \mathbb{N}_0}, \mathbb{N}_0 \in \mathbb{N}\) converges \(Q\)-linearly to \(w^*\) around its neighborhood. As for \(0 < \alpha < 1\), from (75), we have:

\[
\mathcal{L}_c^{k+1} - \mathcal{L}_c^* \\
\leq \frac{1}{\alpha c} ||\nu^k - \nu^{k+1}||^2 + c\lambda_B^2 ||q^k - q^{k+1}||^2 \\
+ \frac{c}{2} ||Ap^{k+1} - Ap^*||^2 + C_G||q^{k+1} - q^*|| \\
\leq \left[\frac{\mu_B^2 L_q^2}{\alpha c} + c\lambda_B^2\right] ||q^k - q^{k+1}||^2 \\
+ \frac{c}{2} ||Ap^{k+1} - Ap^*||^2 + C_G||q^{k+1} - q^*|| \\
\leq \left[\frac{\mu_B^2 L_q^2}{\alpha c} + c\lambda_B^2\right] ||q^k - q^{k+1}||^2 \\
+ C^* ||Ap^{k+1} - Ap^*||^2 + ||q^{k+1} - q^*||^2 \\
+ ||\nu^{k+1} - \nu^*||^2,
\]

where \(2C_G := \left[\sigma_G^2 - c\lambda_B^2(1/\mu_B^2 + 4 - 2\alpha)\right]^{1/2}\), \(C^* := \max\{C_G, c\alpha - 1/2\}\) and the second inequality is due to the \(L_q\)-smoothness of the sub-objective function \(G\). Then using Lemma 8, we have:

\[
\mathcal{L}_c^k - \mathcal{L}_c^* - (\mathcal{L}_c^{k+1} - \mathcal{L}_c^*) \\
\geq \left[\frac{c}{\mu_B^2} \left(1 - \frac{1}{\alpha}\right) - \sigma_G \frac{\mu_B^2 L_q^2}{c\alpha}\right] ||q^k - q^{k+1}||^2 \\
= K_G||q^k - q^{k+1}||^2,
\]
where by assumption $K_G > 0$. Combining the above two results, defining the constants $C_q := \mu_B^2 L_q^2/(\alpha c) + c\lambda^2/2$, $K^* := \max\{C_q/K_G, C^*_q\}$, and denote $\Delta^k := L^k_c - L^*_c$, we have:

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

Then by Lemma 13, 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.

**APPENDIX R**

**PROOF OF LEMMA 18**

By construction, since $L_c$ is solved with Solver II, from (71) in Lemma 16, we have:

$$L^{k+1}_c - L^*_c \leq \frac{1}{\alpha c} \|\nu^k - \nu^{k+1}\|^2 + \frac{\sigma_G}{2} \|q^{k+1} - q^*\|^2 + \frac{c}{2} \|Bq^{k+1} - Bq^*\|^2 + \frac{cB}{1-\alpha} \|Bq^k - Bq^{k+1}\|^2 + c\left(1 - \frac{1}{\alpha}\right) \|Bq^k - Bq^k_{k+1}\|^2.$$  

Then applying the Cauchy-Schwarz inequality $\|u + v\|^2 \leq 2(\|u\|^2 + \|v\|^2)$ on the term $\|Bq^k - Bq^*\|^2$, we have:

$$L^{k+1}_c - L^*_c \leq \frac{1}{\alpha c} \|\nu^k - \nu^{k+1}\|^2 + \frac{\sigma_G}{2} \|q^{k+1} - q^*\|^2 + \frac{c}{2} \|Bq^{k+1} - Bq^*\|^2 + \frac{cB}{1-\alpha} \|Bq^k - Bq^{k+1}\|^2 + c\left(1 - \frac{1}{\alpha}\right) \|Bq^k - Bq^k_{k+1}\|^2,$$

where the last inequality is due to the fact that $\alpha + 1/\alpha \geq 2$. Then, by defining $W_1 := \max\{1/(\alpha c), \lambda^2\}$ and $W_G := \max\{\sigma_G/2, \lambda^2\}/(2-\alpha)$, we have:

$$L^{k+1}_c - L^*_c \leq W_1 \left(\|\nu^k - \nu^{k+1}\|^2 + \|q^{k+1} - q^*\|^2\right) + W_G \left(\|q^{k+1} - q^*\|^2 + \|p^{k+1} - p^*\|^2\right) \leq W_1 \left(\|\nu^k - \nu^{k+1}\|^2 + \|q^{k+1} - q^*\|^2\right) + W_G \left(\|w^{k+1} - w^*\|^2\right),$$

where $w^k := (p^k, q^k, w^k)$ denotes the collective point at step $k$. On the other hand, for the lower bound of $\|\nabla L^{k+1}_c\|$, from (73) we have:

$$\|\nabla L^{k+1}_c\|^2 \geq \mu^2_A \left[\|\nu^k - \nu^{k+1}\|^2 + c^2 \|Bq^k - Bq^{k+1}\|^2 - 2cL_q \|q^k - q^{k+1}\|^2\right].$$

Then from (72), we have:

$$L^{k+1}_c - L^*_c \leq W_1 \left(\|\nu^k - \nu^{k+1}\|^2 + W_G \left(\|q^{k+1} - q^*\|^2 + \|p^{k+1} - p^*\|^2\right)\right) \leq W_1 \left(\|\nu^k - \nu^{k+1}\|^2 + \|q^{k+1} - q^*\|^2\right) + W_G \left(\|w^{k+1} - w^*\|^2\right),$$

where $W^* := \max\{W_1/W_2, W_G\}$; the last inequality is due to Lemma 12. Finally, by taking square root of both sides of the above inequality, we prove that the Lojasiewicz exponent $\theta = 1/2$.

**APPENDIX S**

**PROOF OF THEOREM 3**

The convergence of the sequence $\{w^k\}_{k \in \mathbb{N}}$ is due to the sufficient decrease lemma (Lemma 10) and Assumption C. Moreover, by Lemma 18, $L_c$ satisfies the KL property with an exponent $\theta = 1/2$. As in (61), we have for some constant $C' > 0$:

$$(L^k_c)^{1-\theta} - (L^k_{c+1})^{1-\theta} \geq C'(1-\theta) \|\nabla L^k_c\|^{-1} \|w^{k+1} - w^k\|^2.$$  

Then from (72), we have:

$$\|\nabla L^k_c\|^2 \leq \lambda^2 \left[\|\nu^k - \nu^{k+1}\|^2 + (2cL_q + c^2\lambda^2) \|q^k - q^{k+1}\|^2\right] + (c^2\lambda_{B_T} + 1) \|Ap^k - Bq^k\|^2.$$  

Recall the following, due to the updating method of Solver II and the identity (36):

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

For $1 \leq \alpha < 2$, substitute the above into (79), we have:

$$\|\nabla L^k_c\|^2 \leq \lambda^2 \left[\|\nu^k - \nu^{k-1}\|^2 + (2cL_q + c^2\lambda^2) \|q^k - q^{k-1}\|^2\right] + (c^2\lambda_{B_T} + 1) \left[\frac{1}{c\alpha} \|\nu^k - \nu^{k-1}\|^2 + (1 - \frac{1}{\alpha}) \|Bq^k - Bq^{k-1}\|^2\right].$$
Since an identity matrix, function with $\sigma$ scaled negative entropy function hence a strongly convex $\sigma$ is satisfied, and we therefore complete the proof.

Define $S^* := \max\{\lambda_2^2 + c^2 \lambda_{BBST} + 1) / (c^2 \alpha)\}$, $\lambda_2^2 c(2L_q + c\lambda_{BB}) + \lambda_2^2 (1 - \alpha)\},$ we have:

$$\|\nabla L_c^k\| \leq S^* \|w^k - w^{k-1}\|.$$  \hspace{1cm} (80)

Then, by Lemma 11 with (62) replaced by (80), the rate of convergence for the case $1 \leq \alpha < 2$ is $Q$-linear. As for $0 < \alpha < 1$, by assumption for some constant $\tau^* > 0$ the following holds due to Lemma 10:

$$L_c^k - L_c^* \geq \tau^* \left( \|p^k - p^{k+1}\|^2 + \|q^k - q^{k+1}\|^2 \right).$$

Additionally, from (71), discarding negative terms, we have:

$$L_c^{k+1} - L_c^* \leq \frac{1}{c\alpha} \|\mu^k - u^{k+1}\|^2 + \frac{c\alpha}{2} \|q^k - q^{k+1}\|^2$$

$$+ \frac{c(2 - \alpha)}{2} \|Bq^k - Bq^*\|^2$$

$$\leq \frac{\mu_{BBST}cL_q^2}{c\alpha} + \frac{c\alpha}{2} \|q^k - q^{k+1}\|^2$$

$$+ c\alpha^2 (2 - \alpha) \|q^k - q^{k+1}\|^2.$$  \hspace{1cm} (81)

where in the last line we apply the Cauchy-Schwarz inequality to $\|Bq^k - Bq^*\|^2$, use the $L_q^k$-smoothness of $G$ and the full row rank assumption of $B$. Define $p_1 := (\mu_{BBST}cL_q^2) / (c\alpha) + \sigma G / 2 + c\alpha^2 (2 - \alpha) > 0$ and $\rho^* := c\alpha^2 (2 - \alpha)$, we have:

$$L_c^{k+1} - L_c^* \leq \rho_1 \|L_c^k - L_c^{k+1}\|^2 + \rho_2 \|u^{k+1} - u^{k}\|^2$$

$$\leq \rho_{max} \left( L_c^k - L_c^{k+1} + \|w^{k+1} - w^k\|^2 \right),$$

where $\rho_{max} := \max\{\rho_1 / \tau^*, \rho^*\}$. Then by Lemma 13, the rate of convergence of the sequence $\{w^k\}_{k>N_0}, N_0 \in \mathbb{N}$ is $R$-linear for $0 < \alpha < 1$. Combining the two results, we conclude that the rate of convergence is locally linear for $0 < \alpha < 2$.

### APPENDIX T

**Proof of Theorem 4**

Due to the $\varepsilon$-inimal assumptions, the Lipschitz smoothness coefficients for $F$ and $G$ are $L_p := 1 / \varepsilon$ and $L_q = 1 / \varepsilon$ respectively. Moreover, by the formulation (23), $F(p)$ is a scaled negative entropy function hence a strongly convex function with $\sigma F = 1 - \gamma > 0$. As for the function $G(q)$, since $p_{z|x} = Q_{z|x}p_{z|x}$ is a strict restriction, from Lemma 3, $G(q)$ is weakly convex w.r.t. the full row rank matrix $B = Q_z$ with the coefficient:

$$\omega_G := 2N_zN_q \varepsilon \gamma > 0,$$

where $\varepsilon$ is defined as in Lemma 3. Lastly, since $A$ is simply an identity matrix, $\lambda_A = \mu_A = 1$. By substituting the above coefficients into Lemma 6 to obtain the smallest penalty coefficient that assures convergence, it is clear that Assumption A is satisfied, and we therefore complete the proof.

### REFERENCES

[1] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” 2000, arXiv:physics/0004057.

[2] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, “Deep variational information bottleneck,” CoRR, vol. abs/1612.00410, 2016. [Online]. Available: http://arxiv.org/abs/1612.00410

[3] N. Ding and P. Sadegh, “A submodularity-based clustering algorithm for the information bottleneck and privacy funnel,” in Proc. IEEE Inf. Theory Workshop (ITW), Aug. 2019, pp. 1–5.
Teng-Hui Huang (Student Member, IEEE) received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 2015, the M.S. degree in communication engineering from the Graduate Institute of Communication Engineering, National Taiwan University in 2017, and the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2022. His research interests include machine learning, optimization, and information theory.

Aly El Gamal (Senior Member, IEEE) received the M.S. degree in mathematics and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 2013 and 2014, respectively. He is currently a Senior Applied Scientist with Amazon.com. He coauthored a Cambridge University Press book on information theory of interference management, developed an award-winning online course on deep learning, and received multiple research and teaching awards from Purdue University, DARPA, and the Indiana Economic Development Corporation. He also led the Purdue finalist teams at both the DARPA Spectrum Collaboration Challenge and the Indy Autonomous Challenge. His research interests include machine learning and information theory.

Hesham El Gamal (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from Cairo University, Cairo, Egypt, and the Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park, MD, USA.

He worked as a member of the faculty with the Department of Electronics and Communication Engineering, The Ohio State University, from 2001 to 2018. He co-founded immobly inc., a venture-backed spinoff from The Ohio State University pioneering the use of AI technology in multimedia delivery, and worked as its Chief Executive Office from 2012 to 2018. He also worked as an Acting Vice President of Research at Nile University, Cairo, Egypt, from 2015 to 2017, and has held visiting appointments at UCLA and Institute Eurecom. He worked as the Chair of the Department of Electrical and Computer Engineering and the Co-Director of the Institute of Cybersecurity and Digital Trust (ICDT), The Ohio State University, from 2018 to 2021. His prior industrial experience spans positions at Alcatel-Lucent and Hughes Network Systems. He currently works as the Engineering Deputy Dean and the Associate Dean of Academic Affairs with The University of Sydney, Australia. He holds key intellectual property in the area of proactive communications, space-time coding/decoding, and graphical code design. He was recognized as one of the top 1% highly cited and most influential researchers by the ISI web of science from 2014 to 2017. He is a fellow of the National Academy of Inventors in 2019, and also a member of the SPACOM Technical Committee. He was a recipient of the OSU Innovator Award in 2013, the OSU Stanley E. Harrison Award in 2008, the National Science Foundation CAREER Award in 2004, the OSU College of Engineering Lumley Research Award in 2002, the OSU Electrical Engineering Department FARMER Young Faculty Development Fund from 2004 to 2008, and the Hughes Network Systems Annual Achievement Award in 2000. He was the Co-Chair of the IEEE GLOBECOM 2008 Communication Theory Symposium and the Co-Chair of the 2010 IEEE Information Theory Workshop. He has served on various editorial boards and conference committees, such as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, an Associate Editor for the IEEE TRANSACTIONS ON MOBILE COMPUTING, and a Guest Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY Special Issue on Cooperative Communications.