Information Bottleneck Methods for Distributed Learning

Parinaz Farajiparvar*, Ahmad Beirami†, and Matthew Nokleby‡

*Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI
†Research Laboratory of Electronics, MIT, Cambridge, MA
‡Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI

Email: {parinaz.farajiparvar, matthew.nokleby}@wayne.edu, berami@mit.edu

Abstract—We study a distributed learning problem in which Alice sends a compressed distillation of a set of training data to Bob, who uses the distilled version to best solve an associated learning problem. We formalize this as a rate-distortion problem in which the training set is the source and Bob’s cross-entropy loss is the distortion measure. We consider this problem for unsupervised learning for batch and sequential data. In the batch data, this problem is equivalent to the information bottleneck (IB), and we show that reduced-complexity versions of standard IB methods solve the associated rate-distortion problem. For the streaming data, we present a new algorithm, which may be of independent interest, that solves the rate-distortion problem for Gaussian sources. Furthermore, to improve the results of the iterative algorithm for sequential data we introduce a two-pass version of this algorithm. Finally, we show the dependency of the rate on the number of samples k required for Gaussian sources to ensure cross-entropy loss that scales optimally with the growth of the training set.

Index Terms—Machine Learning, Rate-distortion function, Information Bottleneck, Distributed Learning, Streaming Data.

I. INTRODUCTION

We consider a distributed learning problem in which Alice obtains a training sequence of k i.i.d. samples drawn from a distribution that belongs to a parametric family. Alice wants to distill the training set to a set of features $T$, which she communicates to Bob using no more than $R$ bits. Bob uses $T$ to estimate the data distribution associated with the learning problem. We measure the quality of Bob’s distribution according to the cross entropy loss, which is ubiquitous in machine learning [1] and closely related to the KL divergence between the learned and true distributions.

This setting induces a rate-distortion problem: For a given bit budget $R$, what is the encoding $T$ of Alice’s training set $X^k$ that minimizes Bob’s cross-entropy loss? We consider this problem for both batch and sequential data, and we show that the ideal strategy is to solve a version of the information bottleneck (IB) problem for an appropriate sufficient statistic of $X^k$ [2].

In this work, the associated rate-distortion problem is equivalent to a special case of the information bottleneck [2], in which one wishes to find a compressed representation, $T$, of an observed random variable $X$ that is maximally “relevant” to a correlated random variable $Y$, as measured by $I(Y; T)$. IB has been applied to clustering and feature extraction [3], [4], and Tishby recently proposed an explanation for the success of deep learning in terms of IB [5]. Extensions of IB to distributed [6], interactive [7], and multi-layer [8] multi-terminal settings have recently been considered. Furthermore, IB was shown to solve distributed learning problems with privacy constraints [9].

In Section III, we show that tailored implementations of IB algorithms proposed in [2], [10] can be used to compute the rate-distortion for discrete and Gaussian data sources. This implementations exploit the fact that using the sufficient statistic reduces the computational/storage complexity of the IB algorithm from exponential in $k$ to polynomial in $k$ and from $kd$-dimensional matrix to $d$-dimensional matrix in discrete and Gaussian sources, respectively. Furthermore, we consider the relationship between the number of samples $k$ and the required compression rate $R$. Indeed, in the data-limited regime in which $k$ is small, a higher $R$ does not impact the distortion significantly.

In Section IV, we consider the encoding of sequential data, where the figure of merit is the total cross-entropy regret. Explicit minimization of the regret turns out to be challenging, so we propose a “greedy” on-line method which gives a tractable approach to encoding Gaussian data. In this method, the agent chooses the encoding considers the cross-entropy regret only at the current time instance, and it is provably suboptimum. To improve the regret performance, we also propose a “two-pass” solution which includes a backwards pass in which the feature encoding is improved by considering the impact on future regret.

Finally, in Section V we draw conclusions and suggest areas for future work.

II. PROBLEM STATEMENT

We consider the unsupervised learning problem for both batch and sequential data, in which data is distributed according to $p(x|\theta)$ and the objective is to minimize the cross entropy of the learned distribution with $p(x|t)$.

A. Batch data

Let $(X, \theta) \in (\mathcal{X}, \Theta)$ be (discrete or continuous) random variables with joint distribution $p(x|\theta)p(\theta)$. The conditional distribution $p(x|\theta)$ represents a parametric family of distributions on $X$, and $p(\theta)$ represents a (known) prior over the family. Alice does not observe $\theta$ directly, but instead observes a set of $k$ i.i.d. samples $X^k := (X_1, \ldots, X_k)$, with $X_i \sim p(x|\theta)$. Alice constructs a distilled representation $T$ of
that an encoder has direct access to \( \theta \) and wishes to construct a compact representation \( T \) to be stored or transmitted to another agent. The representation \( T \) is used to construct an approximation \( p(x|T) \) of the parametric distribution, which can be used to predict hypothetical test points \( X \sim p(x|\theta) \). This gives rise to the Markov chain \( X - \theta - X^k - T \), where we emphasize that \( X^k \) is the training set, and \( X \) is a hypothetical test point conditionally independent of \( X^k \) given \( \theta \). Here, we suppose

\[
(X, X^k) \sim p(x|\theta) \quad \text{Alice} \quad T \quad \text{Bob} \quad p(x|T)
\]

Fig. 1: Batch data transmission model.

use of the iterative IB algorithm from [2] requires iteration over all \( |X|^k \) possible training sets, which is unmanageable in practice; in Section III we show how to reduce the computational and storage burden. Further, when \( X \) and \( \theta \) are jointly Gaussian, we specialize the results from [10] to derive a simple, closed-form expression for \( D_k^B(R) \).

The trade-off described by \( D_k^B(R) \) improves with larger \( k \). The rate-distortion curve may be poor for small \( k \), while \( \lim_{k \to \infty} D_k^B(R) \) approaches the special case in which Alice has direct access to \( \theta \). If \( k \) is small, there may be little point in using many bits to describe \( X^k \).

Fig. 2: Rate-Distortion curve, which shows the outer bounds on the rate and distortion.

B. Sequential data

Next, we consider sequential data, where again \( (X, \theta) \in (X, \Theta) \) be (continuous) random variables with joint distribution \( p(x|\theta)p(\theta) \). In this case, instead of observing a set of \( k \) i.i.d. samples, Alice observes the data one-by-one in each round of the sequential data transmission, where samples are drawn from \( p(x|\theta) \). After each round \( l \), Alice constructs a distilled representation \( T_l \) of the training set that she observes up to \( l \)-th round where \( 1 \leq l \leq k \) and transmits it to Bob (Figure 3). Then, Bob uses \( T^l \) to construct the distribution \( p(x|T^l) \), where \( T^l := \{T_1, T_2, \ldots, T_l\} \). In this set up, the

Fig. 3: Sequential data transmission model.
1) Comprehensive solution: In this solution, we consider the global design of the features $T_l$ while ensuring that $T_l$ respects causality. In this case, $I(X^l;T_l|T^{-l})$ represents the number of bits required to describe $T_l$ given that the features $T^{-l}$ have already been sent. We take the overall distortion to be the sum regret, or the suffered cross-entropy loss summed over all $k$ samples. Then, distortion-rate function is defined as the minimum distortion that we achieve when the bit budget of each round $1 \leq l \leq k$ is less than $R_l$:

$$D_{Sc}^k(R) := \min_{p(t_1|x^1),\ldots,p(t_k|x^k):I(X^l;T_l|T^{-l}) \leq R_l} \sum_{l=1}^k H(X|T^l).$$

Based on $I(X^l;T_l) = H(X) - H(X|T^l)$, we can translate the distortion-rate function to the following variational problem:

$$\max_{p(t_1|x^1),\ldots,p(t_k|x^k)} \mathcal{L} = \sum_{l=1}^k I(X^l;T_l) - \alpha_l I(X^l;T_l|T^{-l})$$

where $T^{-l} := \{T_1, T_2, \ldots, T_{l-1}\}$, $I(X^l;T_l|T^{-l})$ determines the rate in each round, and $\alpha_1, \ldots, \alpha_k$ shows the trade-off between the average distortion and the rate of each round. Unfortunately, this problem results in a challenging joint optimization problem over the features $T_l$, and even for Gaussian data, finding a closed-form solution is challenging.

2) Online solution: To find a tractable solution, we present an on-line approach. Instead of find the global solution to the encodings $T_l$, we optimize each term in the objective function one-by-one, without regard for future terms. Therefore, in the $l$th round, we have the distortion-rate function

$$\mathcal{D}_{So}^l(R) := \min_{p(t_l|x^l):I(T_l;X^l|T^{-l}) \leq R_l} H(X|T^l),$$

(1)

where the stochastic mapping $p(t_l|x^l)$ gives a “soft” description of $T_l$ in each round.

In this setup, we define $I(X^l;T_l|T^{-l}) = H(X|T^{-l}) - H(X|T^l)$ as the average distortion. This translates the distortion-rate problem to the following minimization problem:

$$\min_{p(t_l|x^l)} \mathcal{L} = I(X^l;T_l|T^{-l}) - \beta_l I(X^l;T_l|T^{-l})$$

where $T_l$ denotes the compressed representation of an observed random variable $X^l$. This formulation is not equivalent to the IB. Instead, the equivalent problem is to minimizing the mutual information $I(X^l;T_l|T^{-l})$ given a constraint on the conditional mutual information $I(X;T_l|T^{-l})$. Consequently, the standard IB algorithms cannot be applied here. In Section IV we develop a new iterative algorithm for computing $\mathcal{D}_{So}^l(R)$ based on the sufficient statistic of the Gaussian distribution.

3) Two-path Solution: In the on-line algorithm presented above, the encoding $T_l$ is chosen supposing that the encoding function for previous features is fixed, and without regard for future features. This results in a strictly suboptimum solution. To improve the solution, we develop a two-pass solution, which adds a backwards pass to the algorithm, taking the future encoding designs as fixed and without regard for previous feature encodings. After carrying out the on-line algorithm above, we optimize the following loss function for the backward path:

$$\min_{p(t_{l-1}|x^l)} \mathcal{L} = I(X_{l-1};T_{l-1}|T^{-(l-1)},T_l) - \beta I(X;T_{l-1}|T^{-(l-1)},T_l)$$

Similar to the online solution, the backward loss function is equivalent to the conditional IB, and we can not obtain the results by standard IB algorithm. As a result, we derive an algorithm that solves the backward path.

III. Batch Data

In this section, we consider the batch data, in which $k$ i.i.d. samples are observed at one time. In this setting, we analyze both discrete and continuous distributions, in terms of the fundamental limits and algorithmic method.

A. Fundamental Limits

To find $D_{B}^k(R)$, we leverage the equivalence between this distortion-rate problem and the information bottleneck over the Markov chain $X - \theta - X^k - T$. Considering $I(X^k;T) = H(X) - H(X|T)$ and solving the distortion-rate problem by Lagrange multiplier translates the distortion-rate problem to finding the conditional distribution $p(t|x^k)$ that solves the problem

$$\min_{p(t|x^k)} \mathcal{L} = I(X^k;T) - \beta I(X;T),$$

(2)

where $I(X^k;T)$ determines the rate, $I(X;T) = H(X) - H(X|T)$ determines the average distortion, and $\beta$ determines the trade-off between the two and dictates which point on the rate-distortion curve the solution will achieve. For discrete sources, one can use the iterative method proposed in [2] to solve (2). This method only guarantees a local optimum, but it performs well in practice. When $\theta$ and $X$ are jointly Gaussian, one can use the results in [10] for the Gaussian information bottleneck, in which the optimum $p(t|x^k)$ is Gaussian and given by a noisy linear compression of the source.

Discrete Source: For $k$-sample with a discrete source, when $X$ and $X^k$ are i.i.d. samples drawn from $p(x|\theta)$, the histogram $H_k$ of $X^k$ is a sufficient statistic for $\theta$, and similar to [2, Theorem 1] the optimum mapping $p(t|x^k)$ that minimize (2) is computed as:

$$q(t|H_k) = \frac{q(t)}{Z(H_k, \beta)} \exp[-\beta D_{KL}[p(x|H_k)||p(x|t)]],$$

where $Z(H_k, \beta)$ is the normalization function and $\beta$ is the Lagrangian multiplier in equation 2.

Gaussian Source: When $X$ and $\theta$ are jointly (multivariate) Gaussian, one can appeal to the Gaussian information bottleneck, [10], where it is shown that the optimum $T$ is a noisy linear projection of the source, which one can find iteratively or in closed form and will be described in the sequel. If $X$ is a $d$-dimensional multivariate Gaussian, however, naive application of this approach requires one to find an $dk \times dk$
projection matrix. Here again, exploitation of the structure of this problem simplifies the result. Without loss of generality, let \( p(x|\theta) = \mathcal{N}(\theta, \Sigma_x) \) and \( p(\theta) = \mathcal{N}(0, \Sigma_\theta) \), where \( \Sigma_x \in \mathbb{R}^{d \times d} \) is the covariance of the data, and \( \Sigma_\theta \) is the covariance of the prior. For \( k \)-sample training set \( X^k \) with a Gaussian source, the sample mean, \( \bar{X}_k = \frac{1}{k} \sum_{i=1}^{k} x_i \) is a sufficient statistic for \( \theta \) and we have the Markov chain \( X - \theta - \bar{X}_k \rightarrow T \), where \( X \) is a hypothetical test point conditionally independent of \( X^k \) given \( \theta \). [10] shows that the IB-optimum compression is \( T = A\bar{X}_k + Z \), where \( Z \) is white Gaussian noise and \( A \) is a matrix whose rows are scaled left eigenvectors of the matrix

\[
M = \frac{\Sigma_x + \Sigma_\theta}{k} + \frac{k-1}{k} \Sigma_\theta - \Sigma_x \Sigma_x^{-1} \Sigma_\theta.
\]

Specifically, let \( \lambda_1, \lambda_2, \ldots \) be the ascending eigenvalues of \( M \) and \( v_1, v_2 \) be the associated left eigenvectors. For \( \beta > 1 \), only the eigenvectors \( v_i \) such that \( \beta_i := (1 - \lambda_i)^{-1} < \beta \) are incorporated into \( A \), and the resulting rate-distortion point is given parametrically by

\[
R(\beta) = \frac{1}{2} \sum_{i=1}^{n(\beta)} \log \left( \frac{1 - \lambda_i}{\beta - 1} \right),
\]

\[
D(\beta) = \frac{1}{2} \sum_{i=1}^{n(\beta)} \log \left( \beta_i \frac{\beta}{\beta - 1} \right) + H(X),
\]

where \( n(\beta) \) is the number of eigenvalues satisfying \( \beta_i < \beta \). For \( \beta \approx 1 \), the rate is small and the distortion is close to the maximum value \( H(X) \); as \( \beta \to \infty \) the rate becomes large and the distortion converges to \( H(X|X^k) \).

The problem setting implies structure on the compression matrix \( A \) beyond what is obvious from above.

**B. Algorithm**

**Discrete Source:** Again, for \( k \)-sample with a discrete source, when \( X \) and \( X^k \) are i.i.d. samples drawn from \( p(x|\theta) \), the histogram \( H_k \) of \( X^k \) is a sufficient statistic for \( \theta \), and we compute \( p(x^k), p(x|X^k) \) and \( p(x, x^k) \) in terms of the histogram. Then, the iterative IB algorithm proposed in [2], can be rewritten

\[
q^{(n)}(t|H_k) = \frac{q^{(n)}(t)}{Z^{(n)}(H_k, \beta)} \exp \left[ -\beta D_{KL} [p(x|H_k)||p(x|t)] \right],
\]

\[
q^{(n+1)}(t) = \sum_{\theta} q^{(n)}(t|H_k) p(H_k),
\]

\[
q^{(n+1)}(x|t) = \frac{1}{q^{(n)}(t)} \sum_{\theta} q^{(n)}(t|H_k) p(H_k, x),
\]

where \( n \) is the iteration index, \( q^{(n)}(t|x^k) \) is the choice for \( p(t|x^k) \) at iteration \( n \), and the other iterated distributions \( q^{(n)}(t), q^{(n)}(x|t) \) are intermediate distributions. The number of terms in the distribution \( q^{(n)}(t|x^k) \) is upper bounded to \( |T|k^{|X|} \) entries by using the histogram of \( X^k \). This distribution must be updated every iteration, and computing the distributions \( q^{(n+1)}(t) \) and \( q^{(n+1)}(x|t) \) requires summing over \( q^{(n)}(t|x^k), p(x^n), \) and \( p(x^k, x) \), the latter two of which are computed based on the histogram and have at most \(|T|k^{|X|} - 1, k^{|X|} - 1 \) entries, respectively. Standard cardinality bounds would suggest that \( |T| = \min \{|\theta|, k^{|X|} - 1\} \) is sufficient to achieve \( R_n^l(D) \). This bound on \( T \) is established along with a fact that \( T \) depends on \( X^k \) only through \( \theta \). As a result, the problem structure allows one to reduce the complexity of the IB algorithm from exponential to polynomial when \(|X| \) is constant, albeit of a potentially large degree. In Figure 4 we plot the rate-distortion curve for a Bernoulli distribution with uniform prior and \( k \in \{4, 10, 20, 50\} \).

**Gaussian Source:** For \( k \)-sample \( X^k \) with the Gaussian distribution, the sample mean \( \bar{X}_k \) is the sufficient statistic for \( \theta \), and the Markov chain is \( X - \theta - \bar{X}_k \rightarrow T \). [10] propose an iterative algorithm to compute the compressed representation \( T^{(n)} = A^{(n)} \bar{X}_k + Z^{(n)} \), where the projection matrix \( A^{(n)} \) and covariance of noise \( Z^{(n)} \sim \mathcal{N}(0, \Sigma_Z^{(n)}) \) is computed as:

\[
\Sigma_Z^{(n+1)} = (\Sigma_{t^{(n)}} x - (\beta - 1) \Sigma_{t^{(n)}})^{-1}
\]

\[
A^{(n+1)} = \beta \Sigma_{t^{(n)}}^{-1} \Sigma_{t^{(n)}} x A^{(n)} (I - \Sigma_{x|X} \Sigma_{t^{(n)} x}^{-1}),
\]

where \( \Sigma_{x|X}^{(n)} \) and \( \Sigma_{t^{(n)}} \) are covariance matrix which is computed based on \( T^{(n)} \) in each iteration. In Figure 5 we plot the rate-distortion curve for a Gaussian source, where \( d = 6 \) and the covariances are drawn elementwise at random from the standard normal distribution and symmetrized. The curve is smooth because the Gaussian information bottleneck provides an exactly optimum solution.

Finally, we show the dependency of the rate on the number of samples \( k \) for Gaussian sources. For the Gaussian sources, as \( k \to \infty \) the gap between the distortion \( h(X|X^k) \) given direct access to the training set and the best case distortion \( h(X|\theta) \) goes to zero; equivalently, the mutual information gap \( I(X; \theta) - I(X; X^k) \) goes to zero. Figure 6 shows the gap between \( h(X|T) \) and \( h(X|X^k) \) for different rate functions. As it demonstrate for \( R = \Omega(\log(k)) \) the gap between
A. Fundamental Limits

Comprehensive Solution: Let’s assume \((X, \theta)\) are jointly Gaussian and we observe a set of \(k\) i.i.d. samples \(X^k = (X_1, X_2, \ldots, X_k)\) and \(X\) is a hypothetical test point conditionally independent of \(X^k\) given \(\theta\) (Figure 3). In sequential data similar to the batch data, sample mean \(\overline{X}_k = \frac{1}{k} \sum_{i=1}^{k} x_i\) is a sufficient statistic for \(\theta\), and the Markov chain is \(X - \theta - \overline{X}_k - T^k\). Then, the computation of \(D_{SC}^k(R)\) is equivalent to the solution of the this problem

\[
\max_{\beta} \mathcal{L} = \sum_{l=1}^{k} I(X; T^l) - \alpha_l I(X_l; T_l | T^{-l})
\]  

where \(I(T^l; X) = h(X) - h(X|T^l)\) determines the average distortion of all rounds and \(I(S_l; T_l | T^{-l})\) determines the rate in each round, \(\alpha_l\) determines the trade-off between the average distortion and rate of each round, and as it is shown in [10] \(T_l\) is a noisy linear projection of the source. The comprehensive solution finds the optimum compression for each round by solving the global problem since this problem is not a convex problem, finding a closed-form solution is highly unlikely. In addition, finding a numerical solution for a Gaussian distribution and even a discrete distribution is computationally expensive as dimension of data \(d\) and the number of samples \(k\) become large. This is because we need to compute \(k(d \times d)\) elements of projection matrix \(A\) for \(k\) samples maximizing the (3).

Online Solution: In the online solution for a streaming data, computation of the \(D_{SC}^k(R)\) equivalent to a conditional information bottleneck, which solves the problem

\[
\min_{A_k} \mathcal{L} = I(\overline{X}_k; T_k | T^{-k}) - \beta I(X; T_k | T^{-k}),
\]  

where \(T^{-k} := \{T_1, T_2, \ldots, T_{k-1}\}\). In this setup \(\beta\) determines the trade-off between rate and relevant information and for \(\beta \approx 1\), the rate of each round \(I(T_k; \overline{X}_k | T^{-k})\) is small and the distortion \(H(X|T^{-k})\) is close to \(H(X|T^{-k})\), when \(\beta \to \infty\) the rate becomes large and the distortion converges to \(H(X|\overline{X}_k)\). In the following theorem, we characterize the optimum \(A_k\):

**Theorem 1.** The optimum projection \(A_k\) that solves (4) for some \(\beta\) satisfies

\[
A_k = \begin{cases} 
[0, \ldots, 0] & 0 < \beta < \beta_{c_1} \\
[\alpha_1 v_1^T, 0, \ldots] & \beta_{c_1} < \beta < \beta_{c_2} \\
& \vdots \\
[\alpha_1 v_1^T, \ldots, \alpha_k v_k^T] & \beta_{c_{k-1}} < \beta < \beta_{c_k}
\end{cases}
\]  

where \(\lambda_1, \lambda_2, \ldots, \lambda_k\) are the ascending eigenvalues of \(M = \Sigma_{\overline{X}_k | T^{-k}} X \Sigma_{\overline{X}_k | T^{-k}}^{-1} v_1, v_2, \ldots, v_k\) are the associated left eigenvectors, and \(\beta_{c_i} = \frac{\lambda_i}{\sum_{j=1}^{k} \lambda_j} \) are critical values for \(\beta\) and

\[
\alpha_1 = \sqrt{\frac{\beta(1 - \lambda_1)}{\lambda_1 (\Sigma_{\overline{X}_k | T^{-k}} v_1)^2}}.
\]
Proof. We first rewrite the loss function in terms of the entropies
\[
\min_{A_k} \mathcal{L} = I(X; T_k|T_k) - I(X; T_k|X, T_k)
\]
By taking derivative of the loss function with respect to the $A$
\[
\frac{\delta \mathcal{L}}{\delta A_k} = (1-\beta)(A_k \Sigma_k|_{T-k} A_k^T + I_d)^{-1} 2A_k \Sigma_k|_{T-k} - \beta (A_k \Sigma_k|_{X,T-k} A_k^T + I_d)^{-1} 2A_k \Sigma_k|_{X,T-k}
\]
In order to obtain minimum of the loss function $\mathcal{L}$, we set
\[
\frac{\delta \mathcal{L}}{\delta A_k} = 0 \text{ then we have:}
\]
\[
\beta^{-1} (A_k \Sigma_k|_{X,T-k} A_k^T + I_d)(A_k \Sigma_k|_{X,T-k} A_k^T + I_d)^{-1} A_k = A_k \Sigma_k|_{X,T-k} \Sigma_k|_{X,T-k}^{-1}
\]
This equation is an eigenvalue problem and $A$ is the eigenvector of $\Sigma_k|_{X,T-k} \Sigma_k|_{X,T-k}^{-1}$. Then we can substitute $A_k = UV$ and $V \Sigma_k|_{X,T-k} \Sigma_k|_{X,T-k}^{-1} U = LV$ similar to [10] and rewrite the (6):
\[
\frac{\beta^{-1}}{\beta} [(U \Sigma_k|_{X,T-k} U^T + I_d)(U \Sigma_k|_{X,T-k} U^T + I_d)^{-1}] U = U L
\]
Considering $\Sigma_k|_{X,T-k} = V^{-1} S V$ and $\Sigma_k|_{X,T-k} = V^{-1} S L V$ and multiplying by $U^{-1}$ from left and $U^{-1}(U \Sigma_k|_{X,T-k} U^T + I_d)^{-1}$ by right we will have:
\[
U U^T = [\beta (I - L) - I] (LS)^{-1}
\]
Therefore, $A_k = UV$, in which $V$ is the eigenvector of the $\Sigma_k|_{X,T-k} \Sigma_k|_{X,T-k}^{-1}$ and $U$ is computed based on (7).

Two-pass Solution: To improve the result of the online solution and make the results closer to the optimum solution, we introduce the two-pass version of the online solution that solves the problem from $(k-1)$-th round to the first round by solving following loss function
\[
\min_{A_{k-1}} \mathcal{L} = I(\overline{X}_{k-1}; T_{k-1}|T^{-(k-1)}, T_k) - \beta I(X; T_{k-1}|T^{-(k-1)}, T_k)
\]
Similar to Theorem 1, we characterize the optimum projection matrix from $(k-1)$-round to the previous round $(k-2)$-round in the following theorem:

**Theorem 2.** The optimum projection matrix $A_{k-1}$ that solves (8)
\[
A_{k-1} = \begin{bmatrix}
0, \ldots, 0 \\
[\alpha_1 v_1^T, 0, \ldots] \\
\vdots \\
[\alpha_k v_k^T, \ldots, \alpha_k v_k^T]
\end{bmatrix}, \quad \beta < \beta_{c_k}
\]
where $\lambda_1, \lambda_2, \ldots$ are the associated left eigenvalues of $K = \Sigma_k|_{T-k}, X \Sigma_k|_{X,T-k}, v_1, v_2, \ldots$ are the associated left eigenvectors, and $\beta = \frac{1}{\lambda_1 L}$ are critical values for $\beta$ and $\alpha_i = \sqrt{\frac{\beta(1 - \lambda_i^{-1})}{\lambda_i(\alpha_i v_i^T \Sigma_k|_{X,T-k} v_i)}}$.

Proof. The proof is similar to that of Theorem 1 and to avoid repetition we do not write the proof. □

**B. Algorithm**

**Comprehensive Solution:** Comprehensive solution finds the optimum solution for a block of data with a known length which is encoded sequentially by considering causality. Since we can not find a closed-form solution in the comprehensive case, we solve this optimization problem for $k = 2$ using numerical optimization method (fminunc function in Matlab) to solve (3) and in some area this function finds the numerical optimization methods (fminunc function in Matlab) to solve (3) and in some area this function finds the local minimum. Therefore, we plot the convex hull of the solution since we know the rate-distortion curve is convex.

Fig. 7: Comparison between online and comprehensive solutions by rate-distortion curve, where $K = 2$ and "X" axis is the total rates and "y" axis is the total distortion.
**Online Solution:** Theorem 1 states that the optimal projection matrix of each round in the online solution consists of eigenvalues of $M$, in order to compute the whole projection matrix $A$ for the streaming data $X^k$ we propose an iterative algorithm as follows:

**Algorithm 1 Online Solution for Streaming data**

1. Initiate the values of $\Sigma_\theta$, $\Sigma_n$, and $K$ number of rounds.
2. for each round, $k = 1 : K$ do
3. $\Sigma_{X_k} \leftarrow \Sigma_n / n + \Sigma_\theta$.
4. $\Sigma_{T-k} \leftarrow A \Sigma_{X_k} A^T + I_d$.
5. $\Sigma_C \leftarrow [(\Sigma_x+(k-1)\Sigma_\theta)A_1, \ldots, (\Sigma_x+(k-1)\Sigma_\theta)A_K/k]$
6. $\Sigma_{X_k|T-k} \leftarrow (\Sigma_{X_k} - \Sigma_C \Sigma^{-1}_{T-k} \Sigma_C^T)$.
7. Compute $M$ and eigenvalues $\lambda_1, \lambda_2, \ldots$ and left eigenvectors $v_1, v_2, \ldots$.
8. Compute the projection matrix of each round, $A_k$ based on $5$.
9. Compute the critical values of $\beta$ for each eigenvalue as $\beta_i = \frac{\lambda_i}{\lambda_{max}}$, and initiate value for $\beta_{size}$.
10. $A = diag(A_1, \ldots, A_K)$
11. end for

where $\beta_{size}$ determines the rank of the projection matrix in each round and depends on the $\beta_i$. In this algorithm, first, for each round (from 1 to K) we find the matrix $M$ and its eigenvalues and eigenvectors. Then, we compute the projection matrix $A_k$ of this round according to $\beta_{size}$ and the eigenvalues and eigenvectors of the matrix $M$ and save it to calculate the $M$ for the next round.

Similar to the batch data, from the sufficiency of the sample mean $X_k$, we see that regardless of $k$, the optimum compression of each round $T_k$ is a $d$-dimensional representation of the training set $X^k$; however, in our algorithm the global projection matrix $A$ is a $dk$-dimensional diagonal matrix operator, which we need it to compute the $\Sigma_{T-k}$. In general, in this problem one can derive the optimum operator from the $d$-dimensional matrix $M$ and compute the projection matrix $A_k$ in each round and store it into the $A = diag(A_1, \ldots, A_K)$.

In Figure 8 we show the sample-distortion curve for a jointly Gaussian distribution, where $h(X|\theta)$ and $k$ represent the distortion function and the number of rounds or the number of samples that we use, receptively. In addition, for a fixed rate, as the number of samples increase we see that the distortion decreases; however, for a sufficiently large number of samples, this reduction becomes negligible. This can be interpreted as the eigenvalue of the $M$ matrix tends to 1.

**Two-pass Solution:** In theorem 2 we propose a two-pass solution, where we encode a block data with known length in each round in contrast to the online solution, which we encode the streaming data. Algorithm 2 computes the optimum projection of each round (assuming other rounds are constant) based on this chain $A_1 - A_2 - \cdots - A_K$ and then updates the optimum projection of each iteration according to the backward chain $A_K - A_{K-1} - \cdots - A_1$. We summarized this procedure in Algorithm 2. In each iteration

**Algorithm 2 Two-pass Algorithm**

1. Initiate the values of $\Sigma_\theta$, $\Sigma_n$, $K$ and $N$ covariance of the $\theta$ and noise, number of rounds and number of iteration that we need to run the two-pass algorithm, respectively.
2. for each round, $n = 1 : N$ do
3. for each round, $k = 1 : K$ do
4. $\Sigma_{X_k} \leftarrow \Sigma_n / n + \Sigma_\theta$.
5. $\Sigma_{T-k} \leftarrow A \Sigma_{X_k} A^T + I_d$.
6. $\Sigma_C \leftarrow [(\Sigma_x+(k-1)\Sigma_\theta)A_1, \ldots, (\Sigma_x+(k-1)\Sigma_\theta)A_K/k]$
7. $\Sigma_{X_k|T-k} \leftarrow (\Sigma_{X_k} - \Sigma_C \Sigma^{-1}_{T-k} \Sigma_C^T)$.
8. Compute the projection matrix of each round, $A_k$ based on $5$.
9. $A = diag(A_1, \ldots, A_K)$
10. end for
11. for $k = K - 1 : 1$ do
12. $\Sigma_{T-k} \leftarrow$ eliminate $k$-th row and $k$-th column of the $\Sigma_{T-k}$.
13. $\Sigma_{X_k|T-k,T_k} \leftarrow (\Sigma_{X_k} - \Sigma_C \Sigma^{-1}_{T-k,T_k} \Sigma_C^T)$.
14. $\Sigma_{X_k|T-k,T_k} \leftarrow (\Sigma_{X_k} - \Sigma_C \Sigma^{-1}_{T-k,T_k} \Sigma_C^T)$.
15. Compute the projection matrix of each round, $A_k$ based on $5$.
16. $A = diag(A_1, \ldots, A_K)$
17. Update $\Sigma_{T-k} \leftarrow A(n)\Sigma_{X_{T-k}}A(n)^T + I_{Kd}$
18. end for
19. end for

of this algorithm, at the first part, we compute the covariance matrix $\Sigma_{T-k}$ and projection matrix $A$. In the second part of the algorithm, we compute $\Sigma_{T-k}$ based on the $\Sigma_{T-k}$ and then compute the projection matrix of each round $A_k$ and update the whole projection matrix $A^{(n)}$, then update the $\Sigma_{T-k}$ and use it for computation of the next round of the data. At the end, we compute the projection matrix for the $n$-th iteration $A^{(n)}$. 

![Number of Samples-Distortion curve for jointly Gaussian distribution (d=10 R={4, 8, 10, 14, 16})](image)
Fig. 9: Rate-Distortion curve for jointly Gaussian distribution with $d = 10$ and $k = \{3, 4\}$ for the online algorithm and $k = \{3, 4\}$ for two-pass algorithm after one iteration.

We show the difference between distortion and its lower bound in terms of the rate, in Figure 9, where we show that the two-pass algorithm has the better performance in compare of the online algorithm. The two-pass algorithm improves the results; however, based on the number of iteration that we want to run the algorithm it needs more computation and time and also it uses the sequential data with the known length, so it is not suitable for the streaming communication.

The rate-distortion curve, illustrated in figure 10, compares the comprehensive and two-pass solutions. In this figure similar to the figure 7 we demonstrate the total distortion versus the total rate that we used. Since in the two-pass solution we add a return path to the online solution, this solution converges to the same value of distortion at the cost of higher rate.

Fig. 10: Rate-distortion curve for both comprehensive and two-pass solutions, where $K = 2$ and x axis is the total rate and y axis is the total distortion.

V. CONCLUSION

We have examined distributed supervised learning from a compressed representation of a $k$-sample batch and streaming training set up to a cross-entropy loss, showing that solving for the distortion-rate function is equivalent to an appropriately modified information bottleneck problem. We derived a greedy method to solve the distortion-rate function for streaming data as well as an algorithm for this problem. Finally, we improved our results for the streaming data by a new two-pass algorithm. A variety of interesting problems remain to be investigated. The first is supervised learning for batch and streaming dataset. For unsupervised learning, linear compression is sufficient for Gaussian sources per [10]; establishing this result for the supervised case, and/or determining the best linear compression for general continuous sources, is a topic for further study. The second is interactive compression of geographically separated training sets. For Gaussian unsupervised sources, results in [7], [12] can be used to establish that a single round of interaction is sufficient for optimality; study of the general case is of interest. The last is the study of the effects of single-shot coding on the rate-distortion function as a function of the data distribution and the number of training samples $k$.

REFERENCES

[1] L.-Y. Deng, “The cross-entropy method: a unified approach to combinatorial optimization, monte-carlo simulation, and machine learning,” 2006.
[2] N. Tishby, F. C. Pereira, and W. Bialek, “The information bottleneck method,” arXiv preprint physics/0004057, 2000.
[3] N. Slonim and N. Tishby, “Document clustering using word clusters via the information bottleneck method,” in Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2000, pp. 208–215.
[4] N. Slonim, G. S. Atwal, G. Tkačik, and W. Bialek, “Information-based clustering,” Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 51, pp. 18 297–18 302, 2005.
[5] N. Tishby and N. Zašlasky, “Deep learning and the information bottleneck principle,” in Information Theory Workshop (ITW), 2015 IEEE, IEEE, 2015, pp. 1–5.
[6] I. E. Aguerri and A. Zaidi, “Distributed information bottleneck method for discrete and gaussian sources,” arXiv preprint arXiv:1709.09082, 2017.
[7] M. Vera, L. R. Vega, and P. Piantanida, “Collaborative representation learning,” arXiv preprint arXiv:1604.01435, 2016.
[8] Q. Yang, P. Piantanida, and D. Gündüz, “The multi-layer information bottleneck problem,” arXiv preprint arXiv:1711.05102, 2017.
[9] B. Moraffah and L. Sankar, “Privacy-guaranteed two-agent interactions using information-theoretic mechanisms,” IEEE Transactions on Information Forensics and Security, 2017.
[10] G. Chechik, A. Globerson, N. Tishby, and Y. Weiss, “Information bottleneck for gaussian variables,” Journal of Machine Learning Research, vol. 6, no. Jan, pp. 165–188, 2005.
[11] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016, http://www.deeplearningbook.org.
[12] B. Moraffah and L. Sankar, “Information-theoretic private interactive mechanism,” in Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on. IEEE, 2015, pp. 911–918.