ECE 6980
Algorithmic and Information-Theoretic Methods in Data Science

Instructor: Jayadev Acharya — Lectures #12, 13
Scribes: Cody Freitag, Ayush Sekhari — 16th, 18th October, 2017

1 Introduction

So far, we've looked a lot at sample complexity and a little bit at time complexity. In these two lectures, we consider communication complexity and its trade offs with sample complexity.

2 Distributed Hide and Seek Problem

We first define the distributed model we consider in this lecture. Then we define the learning problem we will study in this distributed setting.

2.1 Distributed Model

There are many possible models to study the tradeoffs of sample complexity and communication complexity for distributed learning. We consider the following model from Shamir [S14].[1]

Definition 1 ((b, n, m) Protocol)**.** Let $M_1, M_2, \ldots, M_m$ be machines with communication to a central server, which can be thought of as a public blackboard. Each machine M_t has access to n i.i.d. data points from some distribution. In some order, each machine M_t computes a function f_t from the data it has seen plus the results of the previous machines, where each function's output must not exceed b bits. Finally, the central server computes a final output using the individual machines outputs. This protocol is summarized as follows:

- For $t = 1, \ldots, m$,
  - M_t samples n i.i.d. data points, denoted as X^t.
  - M_t sends $W^t = f_t(W^1, \ldots, W^{t-1}, X^t)$ to the server, not exceeding b bits.
- The server outputs $W = f(W^1, \ldots, W^t)$.

For a particular problem $\mathbb{P}$, we want to figure out the sample complexity of a (b, n, m) protocol that solves $\mathbb{P}$. Essentially, we want to understand the tradeoff between each machine's communication capacity, b, and the total samples needed, mn. If b is unbounded and there is just 1 machine, this corresponds exactly to the standard notion of sample complexity.

For simplicity, we consider the case where $n = 1$, where the number of samples equals the number of machines. This case is at least as hard as the problem for general n when fixing the total number of samples mn. This is because a single machine could simulate n machines solving the $n = 1$ problem for fixed b. We will see how the number of machines, m, depends on the maximum communication, b, and other parameters for this problem.

[1] In fact, all results in these two lectures are adopted from [S14].

2.2 Hide and Seek Problem

At a high level, the Hide and Seek problem is to determine which coordinate is biased in a simple multi-dimensional distribution. The parameters of interest are the dimension d and the bias ρ. We define this formally as follows.

Definition 2 (Hide and Seek Problem)**.** For $j = 1, \ldots, d$, let $Pr_j(\cdot)$ be the distribution on $\{\pm 1\}^d$ such that

- $Pr_j(x_1, \ldots, x_d) = Pr_j(x_1) \cdot \ldots \cdot Pr_j(x_d)$, and
- $Pr_j(x_i = 1) = \begin{cases} 1/2 & \text{if } i \neq j \\ 1/2 + \rho & \text{if } i = j \end{cases}$.

Suppose j is unknown and chosen uniformly at random from $\{1, \ldots, d\}$. Output j given sample access to Pr_j.

We'll first analyze the sample complexity for Hide and Seek when there is no constraint on b. Then we'll give upper and lower bounds on the sample complexity when b is constrained.

3 No Memory Constraint

We first consider the standard notion of sample complexity for the Hide and Seek problem. In other words, how many samples are needed when there is no constraint on b? We claim this if there is no constraint on b, that is equivalent to a $(b, 1, m)$ protocol since each machine can process its data point and send its entire memory to the central server.

If $b \geq \Omega(d \log \frac{1}{\rho})$, we can store and transmit an empirical estimate for each coordinate with $\text{poly}(\rho)$ additive accuracy. To see why the $\log \frac{1}{\rho}$ factor is necessary, consider a case where we need accuracy in each coordinate within 0.1. Then, we would need to store at least $\log \frac{1}{0.1} = \log 10$ bits per coordinate. The empirical estimate needs to be accurate in each coordinate within $\text{poly}(\rho)$, so we have to use $\Omega(\log \frac{1}{\rho})$ bits per coordinate.

We show that if each machine updates an empirical estimate of all samples seen so far, $m = O(\frac{\log d}{\rho^2})$ machines suffice. We then show that this is tight with a lower bound for the standard sample complexity of the Hide and Seek Problem.

3.1 Upper Bound

Theorem 1 (Unconstrained Upper Bound)**.** *For any $b \geq \Omega(d \log \frac{1}{\rho})$, the protocol where each machine updates an empirical estimate of all samples seen so far is a $(b, 1, m)$ protocol for the Hide and Seek problem for $m = O(\frac{\log d}{\rho^2})$.*

Proof. We show that by estimating the empirical bias of each coordinate with $O(\frac{\log d}{\rho^2})$ samples, we can figure out j with at least $9/10$ probability. Let Z_i be the bias at coordinate i. It is easy to see that

$$\mathrm{E}[Z_i] = \begin{cases} 0 & \text{if } i \neq j \\ 2\rho & \text{if } i = j \end{cases}$$

Therefore we need a sufficient number of samples such that $\mathrm{E}[Z_i] < \rho$ for all $i \neq j$ and $\mathrm{E}[Z_j] > \rho$ with at least $9/10$ probability.

Since each coordinate is a Bernoulli distribution, we can estimate each coordinate within an additive factor of ρ after $O(1/\rho^2)$ samples with 9/10 probability. However, in order for every coordinate to simultaneously be right with 9/10 probability, we need the error probability of each individual coordinate to be at most $O(1/d)$. We can do this by repeating $O(\log d)$ times and taking the majority answer. Thus, $O(\frac{\log d}{\rho^2})$ samples suffice to determine j with at least 9/10 probability. □

3.2 Lower Bound

Theorem 2 (Unconstrained Lower Bound)**.** *The Hide and Seek problem has sample complexity* $\Omega(\frac{\log d}{\rho^2})$.

Proof. We'll start by arguing why the $\log d$ factor is necessary. Suppose $\rho = 1/2$. Then we can rule out any coordinate that ever equals -1. For each data point, every non-biased coordinate will be -1 with probability 1/2, so after $O(\log d)$ rounds, we will rule all but the biased coordinate.

We next use an information theoretic argument to show the lower bound for general ρ. First we claim for a data point x, the coordinate x_j gives $O(\rho^2)$ information about j. This is because distinguishing $Bern(1/2)$ from $Bern(1/2 + \rho)$, which is 1 bit of information, requires $\Theta(1/\rho^2)$ samples, so each sample must gives at most $O(\rho^2)$ bits of information. Furthermore, no other coordinate gives any information about j. Then because j was chosen uniformly from $[d]$, we need to get at least $\Omega(\log d)$ bits of information to determine j with at least 9/10 probability. Thus, the sample complexity is at least $\Omega(\frac{\log d}{\rho^2})$. □

This theorem gives the following corollary in the language of (b, n, m) protocols.

Corollary 1. *Any* (b, n, m) *protocol for the Hide and Seek problem requires* $mn \geq \Omega(\frac{\log d}{\rho^2})$.

4 Memory Constraint

We just saw if b is sufficiently large, then the sample complexity of any $(b, 1, m)$ protocol for Hide and Seek is $\Theta(\frac{\log d}{\rho^2})$. If b isn't sufficiently large, we will possibly need more machines to solve the problem.

In this section, we give an upper bound on the number of machines for a $(b, 1, m)$ protocol that requires no knowledge of past messages. We then give a lower bound even assuming machines can use past messages. Furthermore, these bounds are tight up to log factors. The bounds also show an exponential in d increase in sample complexity relative to the non-memory constrained setting for small enough values of b.

4.1 Upper Bound

Theorem 3 (Constrained Upper Bound)**.** *For any* $b \leq O(d \log \frac{1}{\rho})$*, there exists a* $(b, 1, m)$ *protocol for the Hide and Seek problem for* $m = O(\frac{d}{b'\rho^2} \log b' \log \frac{d}{b'})$ *where* $b' = \Theta(\frac{b}{\log(1/\rho)})$.

Proof. Let $b' = \Theta(\frac{b}{\log(1/\rho)})$ be the maximum number of coordinates you accurately store with an empirical estimator. In the protocol, each machine will store an empirical estimate for b' coordinates. By Theorem 1, it will take $O(\frac{\log b'}{\rho^2})$ machines, each using one sample, for the central server to determine if j is in those b' coordinates with at least 9/10 probability. We can repeat

this for all $\frac{d}{b'}$ different chunks of size b'. Thus, with $O(\frac{d \log b'}{b' \rho^2})$ samples we can make sure the central server can determine with at least $9/10$ probability whether or not each chunk contains j. Lastly, we repeat this $O(\log \frac{d}{b'})$ times to amplify the failure probability for each of the $\frac{d}{b'}$ chunks to $O(\frac{1}{d/b'})$. Taking a union bound, this implies that all chunks are correct with at least $9/10$ probability. Thus the number of machines required for this approach is $m = O(\frac{d}{b' \rho^2} \log b' \log \frac{d}{b'}) = O(\frac{d \log(1/\rho)}{b \rho^2} \log \frac{b}{\log(1/\rho)} \log \frac{d \log(1/\rho)}{b})$. □

4.2 Lower Bound

Theorem 4 (Constrained Lower bound)**.** *Any $(b, 1, m)$ protocol for the Hide and Seek problem requires $m \geq \Omega(\frac{d}{b\rho^2})$.*

We note that this lower bound applies even with the knowledge of past messages, which our upper bound from the previous section did not make use of. However, the upper bound is still tight up to log factors.

At a high level, we will prove the theorem by comparing upper and lower bounds on the sum of KL Divergence over all biased distributions Pr_j with an unbiased distribution, which we will call Pr_0. These bounds are information theoretic and will imply the lower bound on m as stated in the theorem.

We first define some useful notation. Then we prove the lower bound on the sum of KL Divergences in Claim 1. Next the upper bound is derived using Claims 2, 3, and 4. Finally, we prove the theorem by comparing the bounds of Claims 1 and 4. In Appendix A, we prove some auxiliary lemmas needed in the following proof in order to clean up the following analysis.

Notation

$W^1, \ldots, W^m$ are the messages generated by the machines in the $(b, 1, m)$ protocol from definition 1. The task is to estimate the biased index j for the Hide and Seek problem as introduced in definition 2.

Recall $Pr_1, \ldots, Pr_d$ are distributions over $\{\pm 1\}^d$, where Pr_j is unbiased everywhere except for the j^th coordinate with bias ρ. We also define a new distribution Pr_0 as

$$Pr_0(x_1, \ldots, x_d) = \frac{1}{2^d},$$

so each coordinate is unbiased and independently distributed over $\{\pm 1\}$.

For simplicity, we will use the notation $\widehat{W}^t$ to denote $(W^1, \ldots, W^t)$. We will exploit notation a bit to think of Pr_j as a function also on $\widehat{W}^t$ for $t \in \{1, \ldots, m\}$. For $i = 1, \ldots, d$, $Pr_i(\widehat{W}^t)$ is a probability distribution corresponding to when the bias is in position i for data points $(X^1, \ldots, X^t)$. Similarly, Pr_0 is a probability distribution over $\widehat{W}^t$ when there is no bias on $(X^1, \ldots, X^t)$.

Lower Bound on Sum of KL Divergences

Claim 1. *Let Pr_i be the biased distribution over $\{\pm 1\}^d$ defined as above and Q be any distribution over $\{\pm 1\}^d$. Suppose after seeing $W^1, \ldots, W^m$ you are able solve the Hide and Seek problem with*

at least $9/10$ *probability, then there exists a constant* c *such that*

$$\sum_{i=1}^{d} D_{\mathrm{KL}}(Pr_i(W^1,\ldots,W^m), Q(W^1,\ldots,W^m)) \geq cd.$$

Proof. Assuming you can solve the Hide and Seek problem, this means after seeing $W^1,\ldots,W^m$ you can figure out j among $[d]$ with at least $9/10$ probability. In particular, this implies you can also distinguish Pr_i from $Pr_{i'}$ for $i \neq i'$ with probability at least $9/10$.

This implies that[2]

$$\Pr(error) \geq \frac{1}{2} - \frac{1}{2} d_{\mathrm{TV}}(Pr_i(W^1,\ldots,W^m), Pr_{i'}(W^1,\ldots,W^m))$$

Thus,

$$d_{\mathrm{TV}}(Pr_i(W^1,\ldots,W^m), Pr_{i'}(W^1,\ldots,W^m)) \geq 8/10.$$

Using triangle inequality on TV-distance, we get that

$$d_{\mathrm{TV}}(Pr_i, Q) + d_{\mathrm{TV}}(Pr_{i'}, Q) \geq 8/10,$$

so for at least one of $\hat{i}$ equal to i or i', $\hat{i}$ must satisfy

$$d_{\mathrm{TV}}(Pr_{\hat{i}}, Q) \geq 4/10.$$

We relate this to the KL Divergence using Pinsker's inequality and conclude

$$D_{KL}(Pr_{\hat{i}}, Q) \geq 2 d^2_{\mathrm{TV}}(Pr_{\hat{i}}, Q) > 3/10.$$

If we do the above analysis for $d/2$ disjoint pairs of indices in $\{1,\ldots,d\}$, we conclude that for at least $d/2$ indices i it must be the case that $D_{KL}(Pr_i, Q) > 4/10$. Thus we get that,

$$\sum_{i=1}^{d} D_{\mathrm{KL}}(Pr_i(W^1,\ldots,W^m), Q(W^1,\ldots,W^m)) > d/2 \cdot (3/10) \geq cd$$

for $c = 3/20$. □

Note that this lower bound for KL Divergence holds for any distribution Q over $\{\pm 1\}^d$. In particular, it must hold for Pr_0 defined above.

Upper Bound on Sums of KL Divergences

We now give an upper bound on the sum of KL Divergence b/w Pr_i and Pr_0 as stated in claim 4. Note that claim 4 is direct application of lemma 4 on claim 2. For the sake of clarity, We will directly use claim 2 for now. An elaborate proof is presented in Appendix A.

Claim 2. *For* Pr_i *and* Pr_0 *as defined before,*

$$D_{\mathrm{KL}}(Pr_i(W^t|\widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1})) \leq c'\rho^2 \mathrm{I}_{X^t\sim Pr_0}(W^t; X_i^t|\widehat{W}^{t-1}).$$

where $I(X,Y|Z)$ *denotes the mutual information between random variables* $X|Z$ *and* $Y|Z$.

[2]Recall that we proved this in Assignment 1, Problem 2(b).

Claim 3.

$$\sum_{i=1}^{d} D_{\mathrm{KL}}(Pr_i(W^t|\widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1})) \leq c'\rho^2 b.$$

Proof. $X_1^t, \ldots, X_d^t$ are independent random variables, so $X_1^t|\widehat{W}^{t-1}, \ldots, X_d^t|\widehat{W}^{t-1}$ are also independent. Then by Lemma 4,

$$\sum_{i=1}^{d} \mathrm{I}(W^t; X_i^t|\widehat{W}^{t-1}) \leq b.$$

For Pr_i and Pr_0 as defined before, using Claim 2,

$$\begin{aligned}\sum_{i=1}^{d} D_{\mathrm{KL}}(Pr_i(W^t|\widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1})) &\leq \sum_{i=1}^{d} c'\rho^2 \mathrm{I}(W^t; X_i^t|\widehat{W}^{t-1}) \\ &\leq c'\rho^2 \sum_{i=1}^{d} \mathrm{I}(W^t; X_i^t|\widehat{W}^{t-1}) \\ &\leq c'\rho^2 b\end{aligned}$$

□

We're now ready to give the desired upper bound on the KL Divergence.

Claim 4.

$$\sum_{i=1}^{d} D_{\mathrm{KL}}(Pr_i(W^1, \ldots, W^m), Pr_0(W^1, \ldots, W^m)) \leq mc'\rho^2 b$$

Proof.

$$\begin{aligned}&\sum_{i=1}^{d} D_{\mathrm{KL}}(Pr_i(W^1, \ldots, W^m), Pr_0(W^1, \ldots, W^m)) \\ &= \sum_{i=1}^{d}\sum_{t=1}^{m} \mathrm{E}_{\widehat{W}^t}[D_{\mathrm{KL}}(Pr_i(W^t|\widehat{W}^t), Pr_0(W^t|\widehat{W}^t))] \\ &\quad \text{(using chain rule for KL Divergence as in Lemma 5)} \\ &= \sum_{t=1}^{m} \mathrm{E}_{\widehat{W}^t}[\sum_{i=1}^{d} D_{\mathrm{KL}}(Pr_i(W^t|\widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1}))] \\ &\leq m \max_{W^1,\ldots,W^t} [\sum_{i=1}^{d} D_{\mathrm{KL}}(Pr_i(W^t|\widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1}))] \\ &\leq mc'\rho^2 b \quad \text{(using Claim 3)}\end{aligned}$$

□

Proof of Constrained Lower Bound

Using Claims 1 and 4 we have,

$$cd \leq \sum_{j=1}^{d} D_{\mathrm{KL}}(Pr_j(W^1, \ldots, W^m), Pr_0(W^1, \ldots, W^m)) \leq mc'\rho^2 b.$$

Thus,

$$m \geq \Omega(\frac{d}{b\rho^2}),$$

which completes the proof.

Thus, we have shown that $\Omega(\frac{d}{b\rho^2})$ samples are required to solve the Hide and Seek problem for any $(b, 1, m)$ protocol, giving a lower bound matching the upper bound of Theorem 3 up to log factors.

5 References

[S14] O. Shamir, Fundamental limits of online and distributed algorithms for statistical learning and estimation, *Advances in Neural Information Processing Systems* 2014. 163–171.

A Appendix

In the following, we state and prove various lemmas used in this lecture. Lastly, we give the proof of Claim 2.

Lemma 1 (Chain Rule for Mutual Information)**.** *Let X and Y be two random variables from marginal distributions P(X), P(Y) and joint distribution P(X, Y). Also, Let I(X;Y) denote the mutual information between X and Y. Then,*

$$\mathrm{I}(X;Y) = D_{\mathrm{KL}}\left(P(X,Y), P(X)P(Y)\right) = \mathrm{E}_x[D_{\mathrm{KL}}(P(Y|X), P(Y))].$$

Lemma 2 (Lemma 4, S14)**.** *Let P and Q be distributions on a discrete set, such that* $\max_x \frac{p(x)}{q(x)} \leq c$*. Then,*

$$D_{\mathrm{KL}}(P,Q) \leq D_{\chi^2}(P,Q) \leq 2cD_{\mathrm{KL}}(P,Q)$$

Lemma 3. *Let* Pr_j*,* Pr_0*,* W^t*,* $\widehat{W}^{t-1}$*, and* X^t *be defined as in Section 3.2. Then,*

$$\begin{aligned} Pr_j(W^t|\hat{W}^{t-1}) &= \sum_{a=\pm 1} Pr_j(W^t|X_j^t = a, \widehat{W}^{t-1}) Pr_j(X_j^t = a|\widehat{W}^{t-1}) \\ &= \sum_{a=\pm 1} Pr_0(W^t|X_j^t = a, \widehat{W}^{t-1}) Pr_j(X_j^t = a|\widehat{W}^{t-1}) \end{aligned}$$

where the last line follows because Pr_j *and* Pr_0 *are identical after fixing the* j^{th} *coordinate.*

Lemma 4. *: Given* $z_1, \ldots, z_d$ *independent random variables and* $W \in [2^b]$ *then*

$$\sum_{i=1}^{d} \mathrm{I}(W; z_i) \leq b.$$

Proof. The claim has a very intuitive formulation that a sample can not give more information about W than the total information contained in W (which is upper bounded by b as W is a b-bit string).

$$\begin{aligned} \sum_{i=1}^{d} \mathrm{I}(W; z_i) &= \sum_{i=1}^{d} H(z_i) - H(z_i|W) && (1) \\ &= H(z_1^d) - \sum_{i=1}^{d} H(z_i|W) && (2) \\ &\leq H(z_1^d) - \sum_{i=1}^{d} H(z_i|W, z_1, \ldots, z_{i-1}) && (3) \\ &= H(z_1^d) - H(z_1^d|W) && (4) \\ &= \mathrm{I}(W, z_1^d) && (5) \\ &= H(W) - H(W|z_1^d) && (6) \\ &\leq b && (7) \end{aligned}$$

Note that lines (1), (5), and (6) come from the definition of mutual information. Line (2) crucially follows from independence. Without independence, the statement isn't true. Line (3) follows since entropy can only decrease when conditioning. Line (4) is the chain rule for conditional entropy. Lastly, line (7) follows since $H(W) = b$ and entropy is always positive.

□

Lemma 5 (Chain Rule for KL Divergence)**.** *For random variables X and Y,*

$$D_{\mathrm{KL}}(P(X,Y), Q(X,Y)) = D_{\mathrm{KL}}(P(X), Q(X)) + \mathrm{E}_x[D_{\mathrm{KL}}(P(Y|X), Q(Y|X))]$$

Proof.

$$\begin{aligned}\sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} &= \sum_{x,y} p(x)p(y|x)[\log \frac{p(x)}{q(x)} + \log \frac{p(y|x)}{q(y|x)}] \\ &= \sum_{x} p(x) \log \frac{p(x)}{q(x)} + \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{q(y|x)}\end{aligned}$$

□

A.1 Proof of Claim 2

Proof. We will prove this by expanding LHS and RHS and equating them.

Expanding the RHS, we get

$$\begin{aligned}\mathrm{I}_{x^t \sim Pr_0}(W^t; X_j^t | \widehat{W}^{t-1}) &= \mathrm{E}_{X^t}\left[D_{\mathrm{KL}}(Pr_0(W^t|X_j^t, \widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1}))\right] \\ &= \sum_{a=\pm 1} Pr_0(X_j^t = a|\widehat{W}^{t-1}) D_{\mathrm{KL}}(Pr_0(W^t|X_j^t, \widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1})) \\ &= \frac{1}{2} \sum_{a=\pm 1} D_{\mathrm{KL}}(Pr_0(W^t|X_j^t, \widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1})).\end{aligned}$$

Expanding the LHS, we get[3]

$$\begin{aligned}&D_{\mathrm{KL}}(Pr_j(W^t|\widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1})) \\ &\leq D_{\chi^2}(Pr_j(W^t|\widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1}))\end{aligned}$$

[3]There might be a better way to manipulate these terms, but we will be sticking to the proof in [S14].

For simplicity, we define $W = W^t|\widehat{W}^{t-1}$ for the following set of inequalities continuing from before.

$$
\begin{aligned}
&= D_{\chi^2}(Pr_j(W), Pr_0(W)) \\
&= \sum_w \frac{(Pr_j(W) - Pr_0(W))^2}{Pr_0(W)} \\
&= \sum_w \frac{(\sum_{a=\pm 1}(Pr_0(W|X_j = a)(1/2 + a\rho) - Pr_0(W|X_j = a))(1/2))^2}{Pr_0(W)} \quad \text{(by Lemma 3)} \\
&= \rho^2 \sum_w \frac{(\sum_{a=\pm 1} aPr_0(W|X_j = a))^2}{Pr_0(W)} \\
&= \rho^2 \sum_w \frac{(\sum_{a=\pm 1}((Pr_0(W|X_j = a) - Pr_0(W))a))^2}{Pr_0(W)} \\
&\leq \rho^2 \sum_w \frac{\sum_{a=\pm 1} 2(Pr_0(W|X_j = a) - Pr_0(W))^2}{Pr_0(W)} \\
&\quad \text{(This follows from the inequality } (a+b)^2 \leq 2(a^2+b^2)) \\
&= 2\rho^2 \sum_{a=\pm 1} \sum_w \frac{(Pr_0(W|X_j = a) - Pr_0(W))^2}{Pr_0(W)} \\
&= 2\rho^2 \sum_{a=\pm 1} D_{\chi^2}(Pr_0(W|X_j = a), Pr_0(W)) \\
&\leq 8\rho^2 \sum_{a=\pm 1} D_{\mathrm{KL}}(Pr_0(W|X_j = a), Pr_0(W)) \quad \text{(by Lemma 2)} \\
&= 16\rho^2 \left(\frac{1}{2} \sum_{a=\pm 1} D_{\mathrm{KL}}(Pr_0(W|X_j = a), Pr_0(W)) \right) \\
&= 16\rho^2 \mathrm{I}_{X^t \sim Pr_0}(W; X_j)
\end{aligned}
$$

The inequality using Lemma 2 follows under the bounds $\frac{Pr_0(W|X_j=a)}{Pr_0(W)} \leq 2$ and $Pr_0(W) = \sum_{a=\pm 1} Pr_0(W|X_j = a)(1/2)$. The last line equates the derived results to the RHS computed for mutual information.

Plugging back in $W = W^t|\widehat{W}^{t-1}$, we finish the proof:

$$D_{\mathrm{KL}}(Pr_j(W^t|\widehat{W}^{t-1}), Pr_0(W^t|\widehat{W}^{t-1})) \leq c'\rho^2 \mathrm{I}_{X^t \sim Pr_0}(W^t; X_j^t|\widehat{W}^{t-1})$$

□