Single-Shot Compression for Hypothesis Testing

Fabrizio Carpi, Siddharth Garg, Elza Erkip

New York University

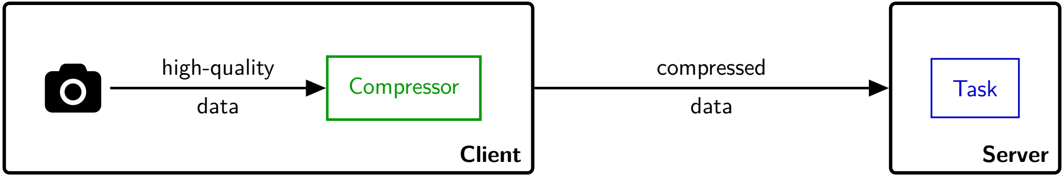
22nd IEEE International Workshop on
Signal Processing Advances in Wireless Communications (SPAWC)

September 27-30, 2021

Outline

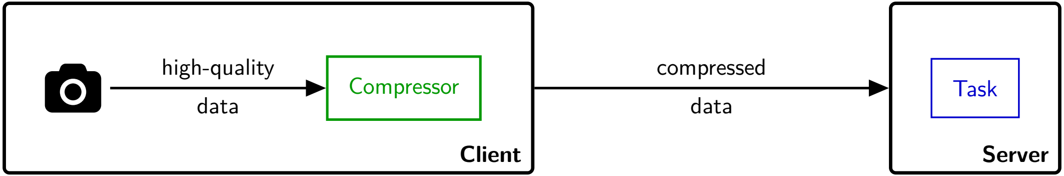
Motivation

A resource constrained **client** offloads costly task-related computations to a remote **server** (edge/cloud computing).



Motivation

A resource constrained **client** offloads costly task-related computations to a remote **server** (edge/cloud computing).



Open question: design **task-aware source coding** schemes which provide *effective* representations of the source data.

In this paper

Assumptions

- **Task**: binary hypothesis testing.
- **Client**: constrained device which cannot perform task locally, does not have memory and can only do simple scalar compression.
- **Server**: hypothesis testing on a block of compressed samples.

In this paper

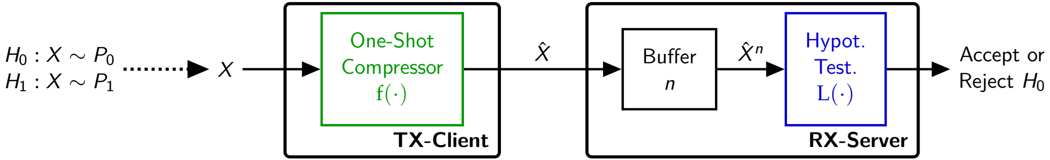
Assumptions

- **Task**: binary hypothesis testing.
- **Client**: constrained device which cannot perform task locally, does not have memory and can only do simple scalar compression.
- **Server**: hypothesis testing on a block of compressed samples.

Our work $\longrightarrow$ single-shot fixed-length compression for hypothesis testing.

- Problem formulation.
- Analyze the error performance.
- Propose a task-oriented compression algorithm for hypothesis testing.

System Model



$X_1, \dots, X_n \sim P_\theta$ are i.i.d. random variables.

Source	**Compressor**	**Hypothesis Testing**
$x \in \mathcal{X} = \{1, \dots, \lvert\mathcal{X}\rvert\}$ $X \sim P_\theta(x),\ \theta \in \{0,1\}$	$\mathrm{f} : \mathcal{X} \to \mathcal{M} = \{1, \dots, M\}$ $\hat{X} = \mathrm{f}(X),\ \hat{X} \sim \widehat{P}_\theta(\hat{X})$	$\mathrm{L}(\hat{X}^n) \underset{\hat{\theta}=1}{\overset{\hat{\theta}=0}{\gtrless}} \log T$

Fixed rate compression $R = \log M$. We consider $M < \lvert\mathcal{X}\rvert$.

Performance Metric

From classical binary hypothesis testing theory[1]:

1. if type-I error $< \epsilon \implies$ type-II error[2] β_n^ϵ decays exponentially in n as $\gamma = -\lim_{n\to\infty} \frac{1}{n} \log \beta_n^\epsilon$.
2. Chernoff-Stein Lemma (without compression): optimal type-II error exponent is $\gamma^\star = D(P_0||P_1)$.

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006. ISBN: 0471241954.

[2] Type-II error: accept H_0 when H_1 is true.

Performance Metric

From classical binary hypothesis testing theory[1]:

1. if type-I error $< \epsilon \implies$ type-II error[2] β_n^ϵ decays exponentially in n as $\gamma = -\lim_{n\to\infty} \frac{1}{n} \log \beta_n^\epsilon$.
2. Chernoff-Stein Lemma (without compression): optimal type-II error exponent is $\gamma^\star = D(P_0 || P_1)$.

Our performance metric $\rightarrow$ type-II error exponent γ.
With compression: the error exponent depends on (f, R): $\gamma_{\mathrm{f}}(R)$.

[1] Cover and Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*.

[2] Type-II error: accept H_0 when H_1 is true.

Performance Metric

From classical binary hypothesis testing theory[1]:

1. if type-I error $< \epsilon \implies$ type-II error[2] β_n^ϵ decays exponentially in n as $\gamma = -\lim_{n\to\infty} \frac{1}{n} \log \beta_n^\epsilon$.
2. Chernoff-Stein Lemma (without compression): optimal type-II error exponent is $\gamma^\star = D(P_0||P_1)$.

Our performance metric $\rightarrow$ type-II error exponent γ.
With compression: the error exponent depends on (f, R): $\gamma_{\mathrm{f}}(R)$.

$\Rightarrow$ We define the **compression penalty**: $\Delta_{\mathrm{f}}(R) = D(P_0||P_1) - \gamma_{\mathrm{f}}(R)$.

[1] Cover and Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*.

[2] Type-II error: accept H_0 when H_1 is true.

Hypothesis Testing on Compressed Variable

Lemma 1

The log-likelihood ratio test on the compressed variables $\hat{X}_i = \mathrm{f}(X_i)$, $i = 1, \ldots, n$, is optimal; the corresponding optimal error exponent is $\gamma_{\mathrm{f}}(R) = D(\widehat{P}_0||\widehat{P}_1)$.

Hence, the **compression penalty** is $\Delta_{\mathrm{f}}(R) = D(P_0||P_1) - D(\widehat{P}_0||\widehat{P}_1)$.

Hypothesis Testing on Compressed Variable

Lemma 1

The log-likelihood ratio test on the compressed variables $\hat{X}_i = \mathrm{f}(X_i)$, $i = 1, \dots, n$, is optimal; the corresponding optimal error exponent is $\gamma_{\mathrm{f}}(R) = D(\widehat{P}_0 || \widehat{P}_1)$.

Hence, the **compression penalty** is $\Delta_{\mathrm{f}}(R) = D(P_0 || P_1) - D(\widehat{P}_0 || \widehat{P}_1)$.

Optimal compressor: $\mathrm{f}^\star = \arg\max_{\mathrm{f}} D(\widehat{P}_0 || \widehat{P}_1) = \arg\min_{\mathrm{f}} \Delta_{\mathrm{f}}$ s.t. $|\mathrm{f}| \leq M$.

NP-hard problem! Optimization over each possible f, which induces a partition of M sets over $\mathcal{X}$.

Compression Penalty: $\Delta_{\mathrm{f}}(R) = D(P_0||P_1) - D(\widehat{P}_0||\widehat{P}_1)$

Proposition 1

Expression for $\Delta_{\mathrm{f}} \geq 0$:

$$\Delta_{\mathrm{f}} = \sum_{\hat{x}=1}^{M} \widehat{P}_0(\hat{x})\, D\Big(P_0(x|\hat{x})\Big|\Big|P_1(x|\hat{x})\Big)$$

where $P_\theta(x|\hat{x}) = \frac{P_\theta(x)}{\widehat{P}_\theta(\hat{x})}\mathbb{1}\{\hat{x} = \mathrm{f}(x)\}$ is the posterior of X given $\hat{X} = \mathrm{f}(X)$.

Compression Penalty: $\Delta_f(R) = D(P_0||P_1) - D(\widehat{P}_0||\widehat{P}_1)$

Proposition 1

Expression for $\Delta_f \geq 0$:

$$\Delta_f = \sum_{\hat{x}=1}^{M} \widehat{P}_0(\hat{x})\, D\Big(P_0(x|\hat{x})\Big|\Big|P_1(x|\hat{x})\Big)$$

where $P_\theta(x|\hat{x}) = \frac{P_\theta(x)}{\widehat{P}_\theta(\hat{x})}\mathbb{1}\{\hat{x} = f(x)\}$ is the posterior of X given $\hat{X} = f(X)$.

Observations:

- ✓ The KL term is zero for one-to-one mappings (or if equal posteriors) $\rightarrow$ only the many-to-one mappings contribute to $\Delta_f(R)$.
- $\rightarrow$ In general, a good task-aware compression strategy combines X values that have similar posteriors.

One-step compression from $|\mathcal{X}|$ to $|\mathcal{X}| - 1$

What is the optimal compressor when reducing the alphabet size by 1?

One-step compression from $|\mathcal{X}|$ to $|\mathcal{X}|-1$

What is the optimal compressor when reducing the alphabet size by 1?

Lemma 2

One-step compression from $|\mathcal{X}|$ to $|\mathcal{X}|-1$: f combines $\{a,b\}\subset\mathcal{X}$ and the others $x\in\mathcal{X}\setminus\{a,b\}$ are one-to-one; i.e., $\mathrm{f}(a)=\mathrm{f}(b)=m\in\mathcal{M}$, $\mathrm{f}(i)=i\in\mathcal{M}\setminus\{m\}$.

Then, the optimal compressor is

$$\mathrm{f}^{\star}=\underset{\{a,b\}\subset\mathcal{X}:\mathrm{f}(a)=\mathrm{f}(b)=m}{\arg\min}\left\{\widehat{P}_0(m)D\Big(P_0(x|m)\Big|\Big|P_1(x|m)\Big)\right\},$$

where $P_\theta(x|m)=\left[\frac{P_\theta(a)}{P_\theta(a)+P_\theta(b)},\frac{P_\theta(b)}{P_\theta(a)+P_\theta(b)}\right]$, $\theta=\{0,1\}$.

Our Proposed Compressor

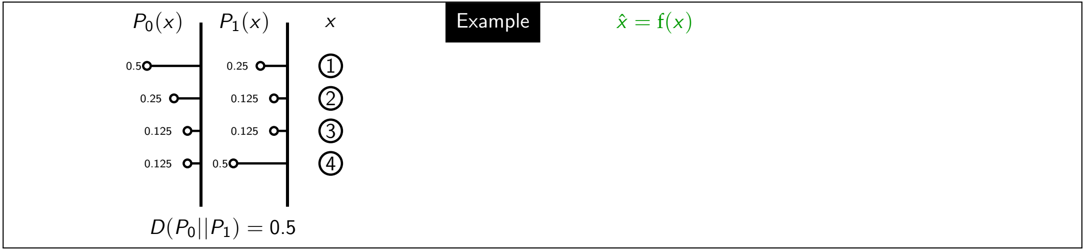
Our "KL-greedy" compressor:

- iteratively reduce the alphabet size by 1 at each step, until the compressed alphabet has size M;
- at each step, combine $\{a, b\}$ which minimize Lemma 2;
- note that this compressor can be determined in polynomial time.

Our Proposed Compressor

Our "KL-greedy" compressor:

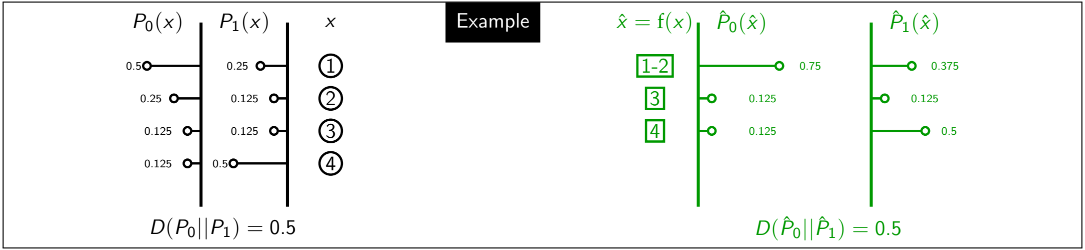
- iteratively reduce the alphabet size by 1 at each step, until the compressed alphabet has size M;
- at each step, combine $\{a, b\}$ which minimize Lemma 2;
- note that this compressor can be determined in polynomial time.



Our Proposed Compressor

Our “KL-greedy” compressor:

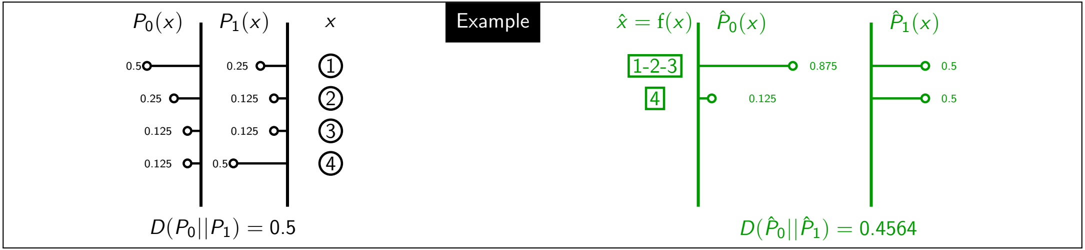
- iteratively reduce the alphabet size by 1 at each step, until the compressed alphabet has size M;
- at each step, combine $\{a, b\}$ which minimize Lemma 2;
- note that this compressor can be determined in polynomial time.



Our Proposed Compressor

Our "KL-greedy" compressor:

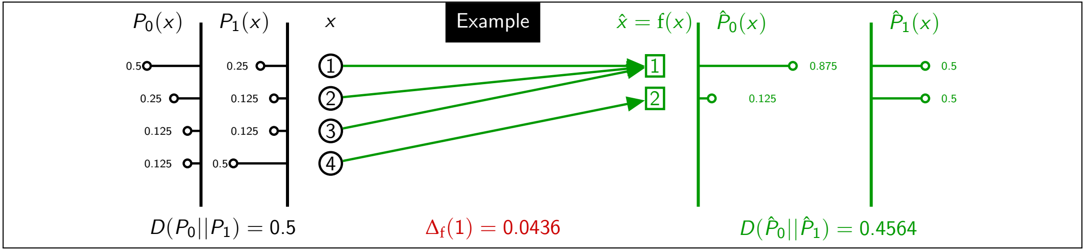
- iteratively reduce the alphabet size by 1 at each step, until the compressed alphabet has size M;
- at each step, combine $\{a, b\}$ which minimize Lemma 2;
- note that this compressor can be determined in polynomial time.



Our Proposed Compressor

Our "KL-greedy" compressor:

- iteratively reduce the alphabet size by 1 at each step, until the compressed alphabet has size M;
- at each step, combine $\{a, b\}$ which minimize Lemma 2;
- note that this compressor can be determined in polynomial time.



Simulation details

P_θ are shifted binomial distributions with different parameters.

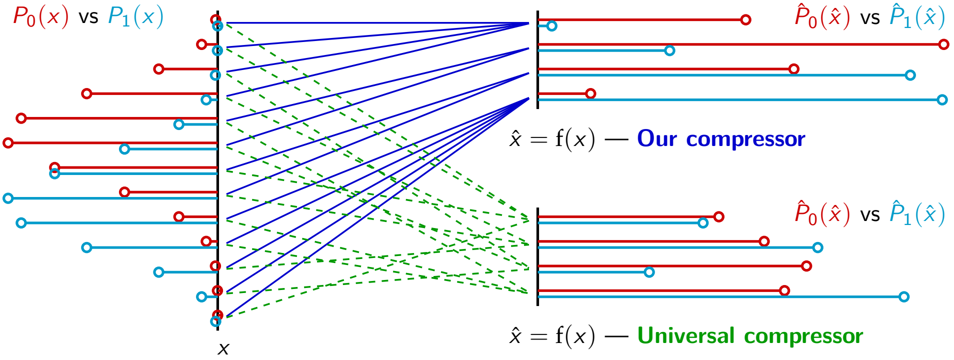
Compare compression penalty Δ_{f} and empirical type-II error rate for:

- optimal compressor $\mathrm{f}^\star$ — when feasible to compute, i.e, small $|\mathcal{X}|$;
- our KL-greedy compressor;
- universal compressor[3] designed for reconstruction under log-loss distortion.

For the empirical type-II error rate, consider a threshold T such that type-I error rate $< \epsilon = 0.05$ for a given compressor at rate M.

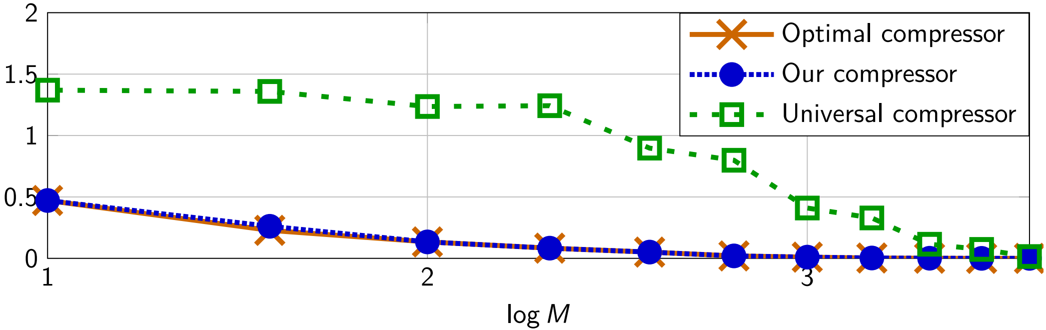
[3] Yanina Shkel, Maxim Raginsky, and Sergio Verdú. "Universal lossy compression under logarithmic loss". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. 2017, pp. 1157–1161. DOI: 10.1109/ISIT.2017.8006710.

Distributions for $M = 4$



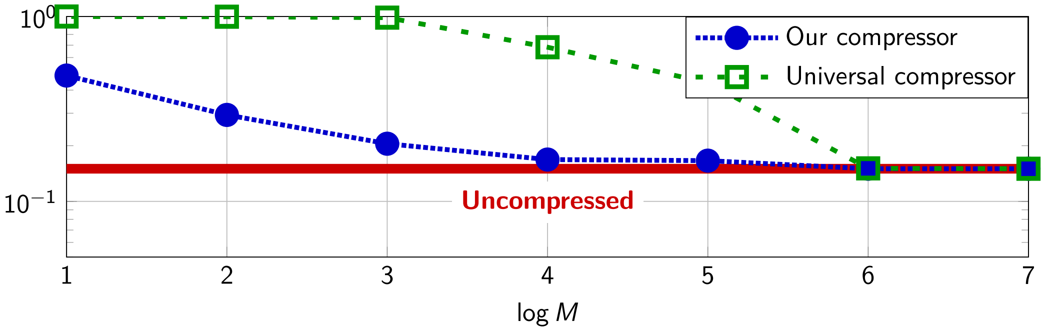
- The compressed KL is larger for our compressor (*divergent* distributions, versus *uniform* in the universal case).
- In our compressor: clustering of source symbols with same *information*.

Compression Penalty $\Delta_f(R)$



- Our compressor performs close to the optimal.
- The compression penalty quickly approaches zero for increasing rate.

Type-II Error Rate for $n = 5$, $\epsilon = 0.05$



- Our compressor achieves error rate close to the uncompressed for increasing rate.

Conclusion

- Formulation for the optimal compressor for hypothesis testing (task-aware).
- Proposed the empirical "KL-greedy" compressor $\rightarrow$ it can be computed in polynomial time and preserves the *useful* information.
- Task-aware compression achieves error rate comparable to the uncompressed case for low rates.

References

Cover, Thomas M. and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006. ISBN: 0471241954.

Shkel, Yanina, Maxim Raginsky, and Sergio Verdú. "Universal lossy compression under logarithmic loss". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. 2017, pp. 1157–1161. DOI: 10.1109/ISIT.2017.8006710.

Thank you! Q&A?

Single-Shot Compression for Hypothesis Testing

Fabrizio Carpi, Siddharth Garg, Elza Erkip

fabrizio.carpi@nyu.edu

NYU TANDON SCHOOL OF ENGINEERING   NYU WIRELESS

22nd IEEE International Workshop on
Signal Processing Advances in Wireless Communications (SPAWC)

September 27-30, 2021

This work was supported in part by NSF–Intel grant #2003182 and NSF grant #1925079.