Information-Distilling Quantizers

Bobak Nazer	Or Ordentlich	Yury Polyanskiy
BU	MIT/BU	MIT

Information Theory and Applications Workshop
February 15, 2017

Focus of this Talk:

- Scalar quantization with the goal of preserving mutual information.
- In particular, what are the fundamental limits of such information-distilling quantizers?
- We focus on the regime where the mutual information to be preserved is itself small.

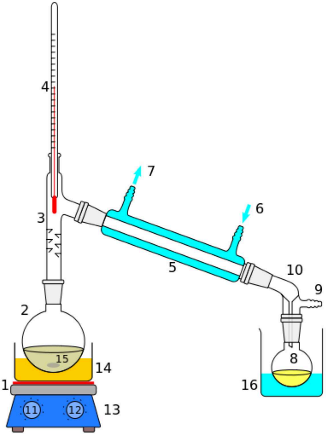
Possible Applications:

- Quantization for low-capacity channels (e.g., continuous to 1 bit output)
- Inference tasks (e.g., clustering while preserving conditional distributions)

Connections:

- Log loss distortion measure
- Information bottleneck
- Polar coding

- Why call it "information-distilling" quantization?
- Better yet, am I even allowed to use the word "distillation"?
- Merriam-Webster defines distillation as
  1. the process of purifying a liquid by successive evaporation and condensation ✗
  2. a process like distillation ✓



Problem Statement

- Let X and Y be random variables with joint distribution P_{XY}.
- Usual notation: Alphabets $\mathcal{X}$, $\mathcal{Y}$ and $[M] \triangleq \{1, 2, \ldots, M\}$.
- Goal: Design an M-ary scalar quantizer f for Y under the objective of maximizing the mutual information between X and $f(Y)$.
- Optimal Quantizer(s): $\underset{f:\mathcal{Y}\to[M]}{\arg\sup}\, I(X; f(Y))$.
- **Our notation:** $I\big(X; [Y]_M\big) \triangleq \sup_{\tilde{Y}\in[Y]_M} I(X; \tilde{Y})$ where $[Y]_M$ is the set of all (deterministic) M-ary quantizations of $\mathcal{Y}$,
$$[Y]_M \triangleq \big\{f(Y) \ : \ f : \mathcal{Y} \to [M]\big\}.$$
- We are mainly concerned with the value of the preserved mutual information (instead of efficient quantizer design algorithms).
- Can show it suffices to consider only deterministic quantizers.

A First Guess

- Take $X \sim \mathrm{Bernoulli}(p)$.
- At a first glance, it might seem that optimal binary quantization suffices to preserve a constant fraction of the mutual information.
- Moreover, it might seem that the MAP quantizer suffices to this end.
- Agrees with our intuition from the AWGN case: the MAP quantizer retains at least $2/\pi \approx 0.637$ fraction of the mutual information.
- For general channels, these intuitions are correct in the large $I(X;Y)$ regime, but not in the small $I(X;Y)$ regime.

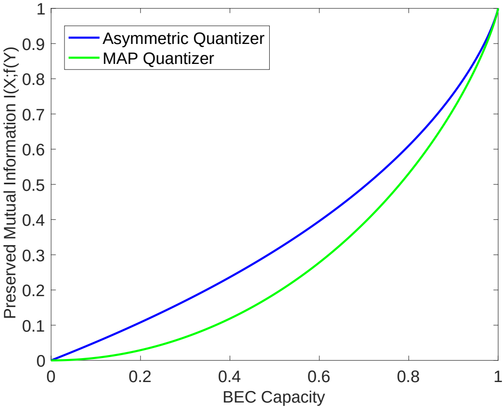
Preview: The BEC

- Consider a standard Binary Erasure Channel (BEC).
- There are only two non-trivial quantizers:

$$f_{\mathsf{MAP}}(y) = \begin{cases} 1 & \text{if } \Pr(X=1|Y=y) > 1/2 \\ 2 & \text{if } \Pr(X=1|Y=y) < 1/2 \\ \mathrm{Bernoulli}(1/2) & \text{if } \Pr(X=1|Y=y) = 1/2 \end{cases}.$$

$$f_Z(y) = \begin{cases} 1 & \text{if } y \in \{1, ?\}, \\ 2 & \text{if } y = 0. \end{cases}$$

Preview: The BEC



- Turns out that, in the small β regime,

$$I(X; f_Z(Y)) = \frac{\beta}{2} h\left(\frac{1-\beta}{2-\beta}\right) + 1 - h\left(\frac{1-\beta}{2-\beta}\right) = \frac{\beta}{2} + o(\beta)$$

$$I(X; f_{\mathsf{MAP}}(Y)) = 1 - h\left(\frac{1-\beta}{2}\right) = \frac{\log e}{2}\beta^2 + o(\beta^2)$$

Connection to Log Loss Distortion Measure

- Log loss distortion for quantizing X: $\mathbb{E}_X\left[\log\left(\frac{1}{q(X)}\right)\right]$
- Assume we would like to quantize Y in order to later make inferences about X. Natural to consider the related distortion measure

$$\mathbb{E}_{XY}\left[\log\frac{P_{X|Y}(X|Y)}{q_Y(X)}\right] = \mathbb{E}_Y\mathbb{E}\left[\log\frac{1}{q_Y(X)}\middle| Y\right] - H(X|Y),$$

- Quantizer f equivalent to selecting partition $\mathcal{S}_1,\ldots,\mathcal{S}_M$ of $\mathcal{Y}$. Let T denote the cell occupied by Y.
- $$\mathbb{E}_Y\mathbb{E}\left[\log\frac{1}{q_Y(X)}\middle| Y\right] = H(X|T) + D(P_{X|T}\,||\,a_T\,|\,P_T) \geq H(X|T)$$

  with equality if and only if $a_t = P_{X|Y\in\mathcal{S}_t}$ for all $t\in[M]$.
- Minimizing $H(X|Y)$ is equivalent to maximizing $I(X;f(Y))$.

Connection to Information Bottleneck

- Recall the information bottleneck tradeoff
(**Tishby - Pereira - Bialek '99, Gilad-Bachrach - Navot - Tishby '03**)

$$\mathrm{IB}_R(P_{XY}) \triangleq \max_{P_{T|Y}\,:\,I(Y;T)\leq R} I(X;T)$$

- Key difference from our formulation is that T can be random and is restricted by $I(Y;T) \leq R$ rather than alphabet size M.
- Studied in machine learning literature.
- Connected to remote source coding.
- Can be interpreted in our context as a single-letter solution as $n \to \infty$ for $P_{X^n Y^n} = P_{X,Y}^n$

$$\lim_{n\to\infty} \frac{1}{n} I(X^n; [Y^n]_{M^n}) = \mathrm{IB}_{\log M}(P_{XY}).$$

- $n = 1$ is of considerable interest since inference is seldom performed in blocks of independent observations.

Worst-Case Information Preservation

- For given P_{XY}, seems difficult to bound $I\big(X;[Y]_M\big)$ in closed-form (and this can be connected to the subset sum problem).
- However, for some special cases, there are polynomial-time algorithms for finding the optimal quantizer. (**Kurkoski** - **Yagi '14**)
- We focus on worst-case bounds in the following sense:
  - Fix input distribution P_X.
  - Fix mutual information β between X and Y.
  - Look for the worst-case channel $P_{Y|X}$.
  - Upper and lower bound resulting $I\big(X;[Y]_M\big)$.
- Formally, we want to characterize the "information-distillation" function:

$$\mathrm{ID}_M(P_X,\beta) \triangleq \inf_{P_{Y|X}\,:\,I(X;Y)\geq\beta} I(X;[Y]_M).$$

Additive Gaps and Connection to Polar Coding

- These quantization questions also appear when constructing efficiently-implementable polar codes. (**Pedarsani - Hassani - Tal - Telatar '11, Tal - Sharov - Vardy '12, Kartowsky - Tal '17**)
- Usual focus is on bounding the additive gap.
- In our notation, **Kartowsky - Tal '17** showed that

$$\mathrm{ID}_M(P_X, \beta) \geq \beta - \nu(|\mathcal{X}|)M^{-2/(|\mathcal{X}|-1)}$$

  for some function ν.

- In the small β regime, the **Kartowsky - Tal '17** quantization approach requires $M = O(\beta^{-1/2})$ to preserve a constant fraction of mutual information.
- In this talk, we show that $M = \Theta(\log(1/\beta))$ to preserve a constant fraction of mutual information for binary-input channels.

Main Result

Theorem (Submitted to ISIT '17)

If $X \sim \mathrm{Bernoulli}(1/2)$*, then*

$$I\big(X;[Y]_M\big) \geq \mathrm{constant} \times \frac{(M-1)\beta}{\log(1/\beta)}.$$

Also, there is a sequence of channels for which this is tight (up to constants).

- A bit more formally: $\mathrm{ID}_M\big(\mathrm{Bernoulli}(1/2), \beta\big) = \Theta\left(\frac{(M-1)\beta}{\log(1/\beta)}\right)$.
- Similar behavior for $\mathrm{Bernoulli}(p)$.
- Explicit constants for upper and lower bounds.

Main Result

Theorem (Submitted to ISIT '17)

If $X \sim \mathrm{Bernoulli}(1/2)$ *and* $I(X;Y) = \beta > 0$, *we have*

$$I(X;[Y]_2) \geq \frac{1}{3e}\frac{\beta}{1+\ln\left(\frac{1}{\beta}\right)}.$$

Furthermore, for any $\eta \in (0,1)$ *and any natural* $M < \frac{12\max\left\{\log\left(\frac{1}{\beta}\right),1\right\}}{(1-\eta)^2}$

$$I(X;[Y]_M) \geq (M-1)\frac{\beta}{\max\{\log\left(1/\beta\right),1\}}\frac{\eta(1-\eta)^2}{12}.$$

Finally, for any $0 < \beta \leq 1$, *there exist distributions* P_{XY} *with* $X \sim \mathrm{Bernoulli}(1/2)$ *and* $I(X;Y) = \beta$, *for which*

$$I(X;[Y]_M) \leq 2M\frac{\beta}{\ln\left(\frac{e\log(e)}{2\beta}\right)},$$

Lemma

For discrete output alphabets $\mathcal{Y}$, $\quad I\big(X;[Y]_M\big) \geq \frac{M-1}{|\mathcal{Y}|} I(X;Y).$

Proof:

- Recall that $I(X;Y) = \sum_{y \in \mathcal{Y}} P_Y(y)\, D(P_{X|Y=y} \| P_X)$.
- Assume $P_Y(1)\, D(P_{X|Y=1} \| P_X) \geq \cdots \geq P_Y(|\mathcal{Y}|)\, D(P_{X|Y=|\mathcal{Y}|} \| P_X)$.
- Set $f(y) = \begin{cases} y & \text{if } y < M, \\ M & \text{otherwise.} \end{cases}$
- Worst case: all $P_Y(y)\, D(P_{X|Y=y} \| P_X)$ values are equal.

Corollary

For natural numbers $K < M$, $\quad I\big(X;[Y]_K\big) \geq \frac{K-1}{M} I\big(X;[Y]_M\big).$

Proof of the Lower Bound

- Define $\alpha_y = \Pr(X = 1|Y = y)$ and $\bar{\alpha} = \mathbb{E}[\alpha_Y]$.
- Also, define $D_y = D\big(P_{X|Y=y} \| P_X\big) = d(\alpha_y \| \bar{\alpha})$.
- Consider the following $M = 2L + 1$ level quantizer:

$$f(y) = \begin{cases} 0 & 0 \le d(\alpha_y \| \bar{\alpha}) < \gamma_1, \\ -\ell & \gamma_\ell \le d(\alpha_y \| \bar{\alpha}) < \gamma_{\ell+1},\ \alpha_y \le \bar{\alpha}, \\ \ell & \gamma_\ell \le d(\alpha_y \| \bar{\alpha}) < \gamma_{\ell+1},\ \alpha_y > \bar{\alpha}. \end{cases}$$

- Follows that

$$\begin{aligned} I(X; f(Y)) &= \sum_{\ell=-L}^{L} \Pr(f(Y) = \ell) D(P_{X|f(Y)=\ell} \| P_X) \\ &\ge \sum_{\ell=1}^{L} \left(\bar{F}(\gamma_\ell) - \bar{F}(\gamma_{\ell+1})\right) \gamma_\ell \\ &= \sum_{\ell=1}^{L} \bar{F}(\gamma_\ell)(\gamma_\ell - \gamma_{\ell-1}), \end{aligned}$$

Proof of the Lower Bound

- Now, set the quantization parameters to

$$\gamma_1 = \frac{I(X;Y)}{L+1} \qquad \theta = \gamma_1^{-1/L} \qquad \gamma_\ell = \gamma_1 \theta^{\ell-1}.$$

  and note that $\gamma_{\ell+1} - \gamma_\ell = \theta(\gamma_\ell - \gamma_{\ell-1})$.

- Let $\bar{F}(\gamma) \triangleq \Pr(D_Y \geq \gamma)$

- We have that

$$\begin{aligned} I(X;Y) = \mathbb{E}[D_Y] = \int_0^{\gamma_{L+1}} \bar{F}(\gamma) d\gamma &= \sum_{\ell=0}^{L} \int_{\gamma_\ell}^{\gamma_{\ell+1}} \bar{F}(\gamma) d\gamma \\ &\leq \sum_{\ell=0}^{L} (\gamma_{\ell+1} - \gamma_\ell) \bar{F}(\gamma_\ell) \\ &= \gamma_1 + \theta \sum_{\ell=1}^{L} (\gamma_\ell - \gamma_{\ell-1}) \bar{F}(\gamma_\ell) \\ &\leq \gamma_1 + \theta I(X; f(Y)) \end{aligned}$$

Proof of the Lower Bound

- Rearranging terms, we have shown that

$$\begin{aligned} I(X; f(Y)) &\geq \big(I(X;Y)\big)^{\frac{L+1}{L}} \frac{L}{(1+L)^{\frac{L+1}{L}}} \\ &\geq \big(I(X;Y)\big)^{\frac{L+1}{L}} \left(1 - \frac{1}{\sqrt{L}}\right) \end{aligned}$$

- We can preserve a constant fraction of mutual information, $I(X; f(Y) \geq \eta I(X;Y)$ with

$$L = \left\lceil \frac{4 \max\left\{ \log\left(\frac{1}{I(X;Y)}\right), 1 \right\}}{(1-\eta)^2} \right\rceil$$

- Recall that $M = 2L + 1$, so $M \leq \left\lceil \frac{12 \max\left\{ \log\left(\frac{1}{I(X;Y)}\right), 1 \right\}}{(1-\eta)^2} \right\rceil$

Counterexample for Upper Bound

- Our upper bound is based on bounding the performance for the following symmetric channel:

$$f_T(t) = \begin{cases} r\delta(t) + \frac{4r}{(1-2t)^3} & 0^- < x \leq \frac{1-\sqrt{r}}{2} \\ 0 & \text{otherwise} \end{cases}$$

- See our preprint for analysis.

A Few Properties

- Data Processing: If $X - Y - V$ form a Markov chain is this order, then $I(X;[V]_M) \leq I(X;[Y]_M)$.

- Convexity: For a fixed P_X, the function $P_{Y|X} \mapsto I(X;[Y]_M)$ is convex.

- Lack of Concavity: For a fixed $P_{Y|X}$, $I\big(X;[Y]_M\big)$ is generally not concave in P_X.

- Monotonicity: The function $\mathrm{ID}_M(P_X, \beta)$ is convex and monotonically nondecreasing in β.

- No Diminishing Returns: The inequality $I(X;[Y]_{M_1 \cdot M_2}) \leq I(X;[Y]_{M_1}) + I(X;[Y]_{M_2})$ is not always satisfied.

Conclusions

- Considered the "information distillation" problem of scalar quantization for preserving mutual information.
- Focused on the regime where the original mutual information β is already quite small.
- For binary input channels, developed upper and lower bounds that are match up to constants.
- Preprint on my website if you are interested.