Functional Representation of Random Variables and Applications

Abbas El Gamal

Stanford University

MIT LIDS, Fall 2018

Based mostly on joint work with Cheuk Ting Li

Functional representation of random variables

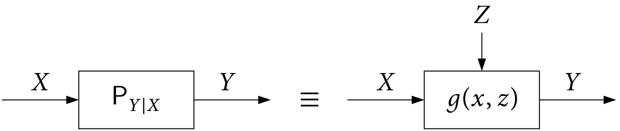
Lemma (see, e.g., EG–Kim (2011))

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$

Functional representation of random variables

Lemma (see, e.g., EG–Kim (2011))

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$



Functional representation of random variables

Lemma (see, e.g., EG–Kim (2011))

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$

- Applications:
  - Broadcast channel (Hajek–Pursley 1979)
  - MAC with cribbing encoders (Willems–van der Meulen 1985)
  - Also see (EG–Kim 2011) for other applications

Functional representation of random variables

Lemma (see, e.g., EG–Kim (2011))

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$

- Applications:
  - Broadcast channel (Hajek–Pursley 1979)
  - MAC with cribbing encoders (Willems–van der Meulen 1985)
  - Also see (EG–Kim 2011) for other applications
  - Entropic causal inference (Kocaoglu–Dimakis–Vishwanath–Hassibi 2017)

Functional representation of random variables

Lemma (see, e.g., EG–Kim (2011))

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$

- Example: B_1, B_2, B_3, B_4 i.i.d. $\mathrm{Bern}(1/2)$, $X = (B_1, B_2, B_3)$, $Y = (B_2, B_3, B_4)$
  - $Z_1 = B_4$, $\quad Y = (B_2, B_3, Z_1)$ $\quad$ (Z_1 part of Y not in X)

Functional representation of random variables

Lemma (see, e.g., EG–Kim (2011))

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$

- Example: B_1, B_2, B_3, B_4 i.i.d. $\mathrm{Bern}(1/2)$, $X = (B_1, B_2, B_3)$, $Y = (B_2, B_3, B_4)$
  - $Z_1 = B_4$, $\quad Y = (B_2, B_3, Z_1)$ $\quad$ (Z_1 part of Y not in X)
  - $Z_2 = B_1 \oplus B_4$, $Y = (B_2, B_3, B_1 \oplus Z_2)$

Functional representation of random variables

Lemma (see, e.g., EG–Kim (2011))

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$

- Example: B_1, B_2, B_3, B_4 i.i.d. $\mathrm{Bern}(1/2)$, $X = (B_1, B_2, B_3)$, $Y = (B_2, B_3, B_4)$
  - $Z_1 = B_4$, $Y = (B_2, B_3, Z_1)$ (Z_1 part of Y not in X)
  - $Z_2 = B_1 \oplus B_4$, $Y = (B_2, B_3, B_1 \oplus Z_2)$
- What Z is most informative about $Y = g(X, Z)$?

$$H(Y|Z_1) = 2 = I(X;Y), \quad H(Y|Z_2) = H(Y) = 3$$

Functional representation of random variables

Lemma (see, e.g., EG–Kim (2011))

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$

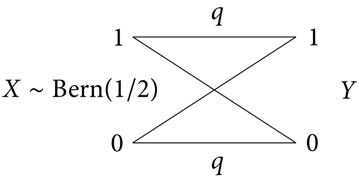
- Example: B_1, B_2, B_3, B_4 i.i.d. $\text{Bern}(1/2)$, $X = (B_1, B_2, B_3)$, $Y = (B_2, B_3, B_4)$
  - $Z_1 = B_4$, $\quad Y = (B_2, B_3, Z_1)$ $\quad$ (Z_1 part of Y not in X)
  - $Z_2 = B_1 \oplus B_4$, $Y = (B_2, B_3, B_1 \oplus Z_2)$
- What Z is most informative about $Y = g(X, Z)$?

$$H(Y|Z_1) = 2 = I(X;Y), \quad H(Y|Z_2) = H(Y) = 3$$

- In general: $H(Y|Z) \geq I(X;Y)$:

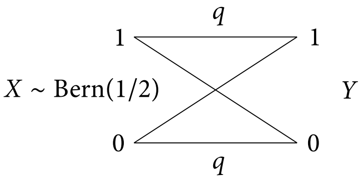
$$\begin{aligned} H(Y|Z) &= I(X;Y|Z) \quad (Y = g(X,Z)) \\ &= I(X;Y,Z) \quad (X \text{ and } Z \text{ independent}) \\ &\geq I(X;Y) \end{aligned}$$

Example (doubly symmetric binary r.v.s)



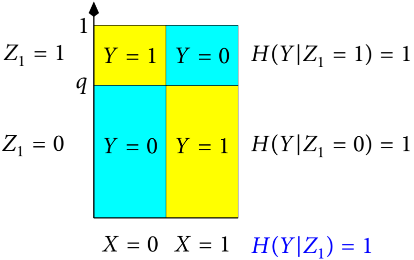
$q = 1 - p,\ p \in (0, 0.5)$

Example (doubly symmetric binary r.v.s)

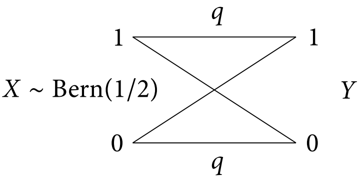


$q = 1 - p,\ p \in (0, 0.5)$

- Let $Z_1 \sim \mathrm{Bern}(p)$ be indep. of X, $Y = X \oplus Z_1$

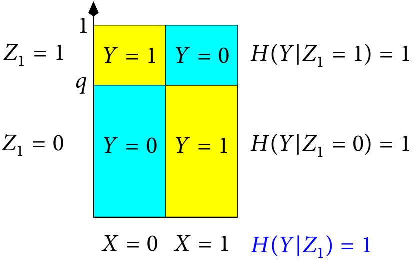


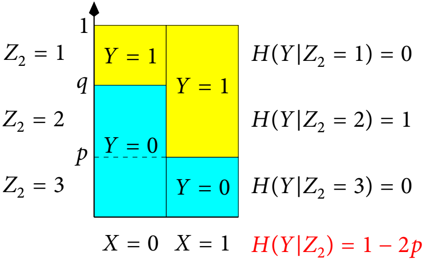
Example (doubly symmetric binary r.v.s)

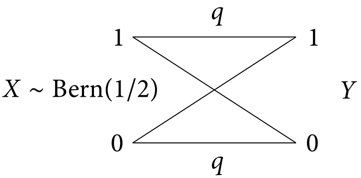


$q = 1 - p,\ p \in (0, 0.5)$

- Let $Z_2 = 1, 2, 3$ w.p. $p, 1 - 2p, p$, respectively, indep. of X

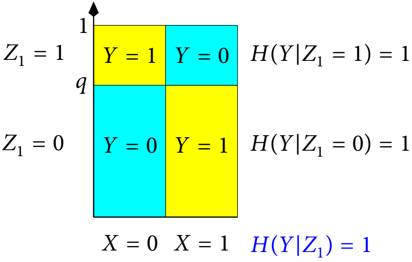


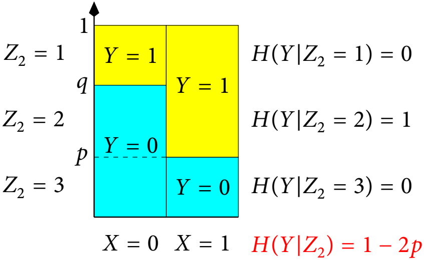




$q = 1 - p, \; p \in (0, 0.5)$

- Can show: $\min_{Z,g} H(Y|Z) = 1 - 2p$, i.e., second construction is optimal
- But $1 - 2p > 1 - H(p) = I(X;Y)$ (cannot always achieve I lower bound)





General upper bound on $H(Y|Z)$

Strong functional representation lemma (SFRL) (Li–EG 2018)

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$, and

$$I(X; Y) \leq H(Y|Z) < I(X; Y) + \log(I(X; Y) + 1) + 4$$

General upper bound on $H(Y|Z)$

Strong functional representation lemma (SFRL) (Li–EG 2018)

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$, and

$$I(X; Y) \leq H(Y|Z) < I(X; Y) + \log(I(X; Y) + 1) + 4$$

- Tighter and more general bound on rate for one-shot channel simulation than in (Harsha et al. 2010), (Braverman–Garg 2014)
- Provides simple achievability results for several coding setups

General upper bound on $H(Y|Z)$

Strong functional representation lemma (SFRL) (Li–EG 2018)

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$, and

$$I(X; Y) \leq H(Y|Z) < I(X; Y) + \log(I(X; Y) + 1) + 4$$

- Tighter and more general bound on rate for one-shot channel simulation than in (Harsha et al. 2010), (Braverman–Garg 2014)
- Provides simple achievability results for several coding setups
- Upper bound can be quite loose, e.g., for binary example with $p = 0.11$,
  - $I(X; Y) = 0.5$, $\min H(Y|Z) = 0.78$, upper bound $= 4.08496$

General upper bound on $H(Y|Z)$

Strong functional representation lemma (SFRL) (Li–EG 2018)

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$, and

$$I(X; Y) \leq H(Y|Z) < I(X; Y) + \log(I(X; Y) + 1) + 4$$

- Tighter and more general bound on rate for one-shot channel simulation than in (Harsha et al. 2010), (Braverman–Garg 2014)
- Provides simple achievability results for several coding setups
- Upper bound can be quite loose, e.g., for binary example with $p = 0.11$,
  - $I(X; Y) = 0.5$, $\min H(Y|Z) = 0.78$, upper bound $= 4.08496$
- For $(X, Y) = (X^n, Y^n)$ i.i.d.: $(1/n)H(Y^n|Z_n) \leq I(X; Y) + O(\log n/n) \approx I(X; Y)$

General upper bound on $H(Y|Z)$

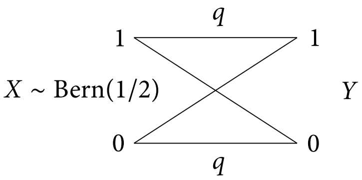
Strong functional representation lemma (SFRL) (Li–EG 2018)

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$, and

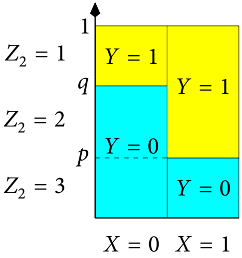
$$I(X;Y) \le H(Y|Z) < I(X;Y) + \log(I(X;Y) + 1) + 4$$

- Tighter and more general bound on rate for one-shot channel simulation than in (Harsha et al. 2010), (Braverman–Garg 2014)
- Provides simple achievability results for several coding setups
- Upper bound can be quite loose, e.g., for binary example with $p = 0.11$,
  - $I(X;Y) = 0.5$, $\min H(Y|Z) = 0.78$, upper bound = 4.08496
- For $(X, Y) = (X^n, Y^n)$ i.i.d.: $(1/n)H(Y^n|Z_n) \le I(X;Y) + O(\log n/n) \approx I(X;Y)$
- There are examples where log term is necessary, SFRL tight within 5 bits

- Recall optimal Z_2 construction for example



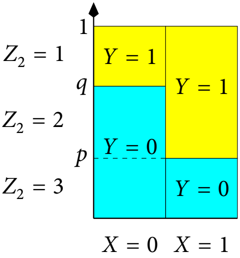
$q = 1 - p,\ p \in (0, 0.5)$

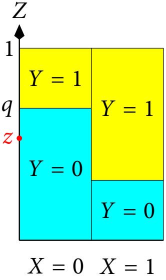


Back to doubly symmetric binary r.v.s example

- Recall optimal Z_2 construction for example
- Can equivalently let $Z \sim \mathrm{Unif}[0,1]$, and:

  For $X = 0$, set $y = 0$ if $\dfrac{z}{q} \le \dfrac{1-z}{p}$; for $X = 1$, set $y = 0$ if $\dfrac{z}{p} \le \dfrac{1-z}{q}$



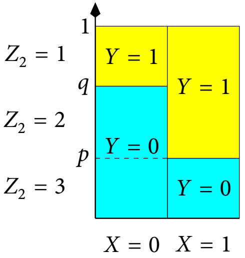


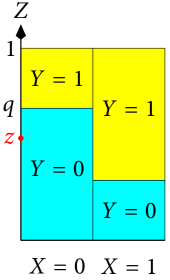
Back to doubly symmetric binary r.v.s example

- Recall optimal Z_2 construction for example
- Can equivalently let $Z \sim \mathrm{Unif}[0,1]$, and:

  For $X = 0$, set $y = 0$ if $\dfrac{z}{q} \le \dfrac{1-z}{p}$; for $X = 1$, set $y = 0$ if $\dfrac{z}{p} \le \dfrac{1-z}{q}$
- In general for $|\mathcal{Y}| = 2$, optimal construction is $Z \sim \mathrm{Unif}[0,1]$ and:

$$y = g(x,z) = \operatorname{argmin}\left\{ \frac{z}{p_{Y|X}(0|x)}, \frac{1-z}{p_{Y|X}(1|x)} \right\}$$





Exponential construction of Z, g

- Let $\mathcal{Y} = \{1, 2, \ldots, l\}$
- Take $Z = (Z_1, Z_2, \ldots, Z_l)$ i.i.d. $\mathrm{Exp}(1)$ r.v.s independent of X, set

$$y = g(x, z) = \operatorname*{argmin}_{y'} \frac{Z_{y'}}{p_{Y|X}(y'|x)}$$

Exponential construction of Z, g

- Let $\mathcal{Y} = \{1, 2, \ldots, l\}$
- Take $Z = (Z_1, Z_2, \ldots, Z_l)$ i.i.d. $\mathrm{Exp}(1)$ r.v.s independent of X, set

$$y = g(x, z) = \operatorname*{argmin}_{y'} \frac{Z_{y'}}{p_{Y|X}(y'|x)}$$

$$\mathsf{P}\{\operatorname*{argmin}_{y'} \mathrm{Exp}(p_{Y|X}(y'|x)) = y\} = p_{Y|X}(y|x) \Rightarrow g(x, Z) \sim p_{Y|X}(.|x)$$

Exponential construction of Z, g

- Let $\mathcal{Y} = \{1, 2, \ldots, l\}$
- Take $Z = (Z_1, Z_2, \ldots, Z_l)$ i.i.d. $\mathrm{Exp}(1)$ r.v.s independent of X, set

$$y = g(x, z) = \operatorname*{argmin}_{y'} \frac{Z_{y'}}{p_{Y|X}(y'|x)}$$

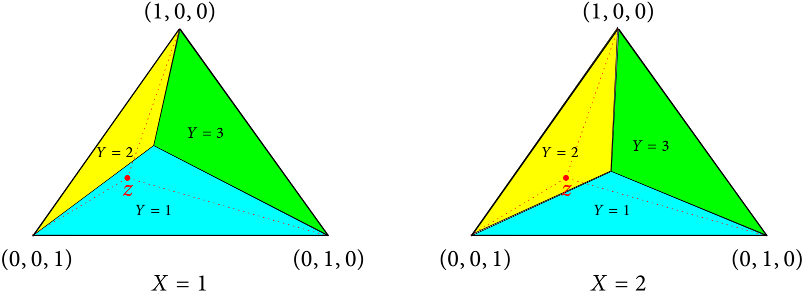
- Let $Z' = (Z'_1, \ldots, Z'_l)$, where $Z'_y = Z_y / \sum_{y'} Z_{y'}$, use above construction of g

Exponential construction of Z, g

- Let $\mathcal{Y} = \{1, 2, \ldots, l\}$
- Take $Z = (Z_1, Z_2, \ldots, Z_l)$ i.i.d. $\mathrm{Exp}(1)$ r.v.s independent of X, set

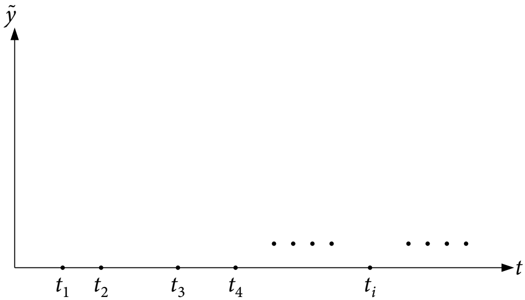
$$y = g(x, z) = \operatorname*{argmin}_{y'} \frac{Z_{y'}}{p_{Y|X}(y'|x)}$$

- Let $Z' = (Z'_1, \ldots, Z'_l)$, where $Z'_y = Z_y / \sum_{y'} Z_{y'}$, use above construction of g
- That is, pick Z uniform over probability simplex in $\mathbb{R}^{|\mathcal{Y}|-1}$



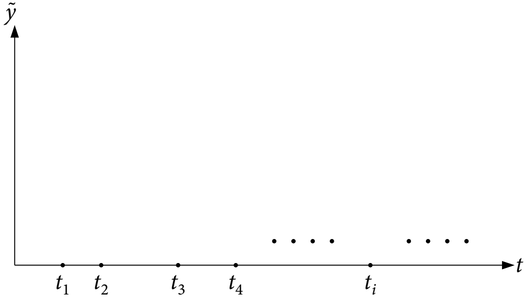
Poisson construction for general Y

- Let $T_1, T_2, \dots$ be arrival times of Poisson process with $\lambda = 1$, indep. of X



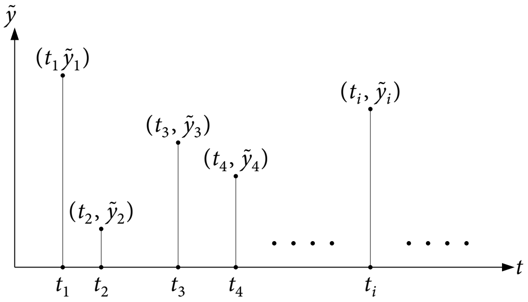
Poisson construction for general Y

- Let $T_1, T_2, \ldots$ be arrival times of Poisson process with $\lambda = 1$, indep. of X
- Let $\tilde{Y}_1, \tilde{Y}_2, \ldots$ be i.i.d. $\sim \mathsf{P}_Y$ indep. of $T_1, T_2, \ldots$ and X



Poisson construction for general Y

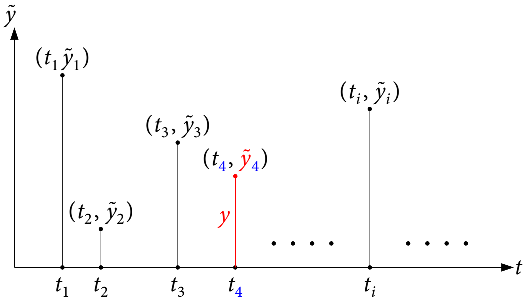
- Let $T_1, T_2, \ldots$ be arrival times of Poisson process with $\lambda = 1$, indep. of X
- Let $\tilde{Y}_1, \tilde{Y}_2, \ldots$ be i.i.d. $\sim \mathsf{P}_Y$ indep. of $T_1, T_2, \ldots$ and X
- Set $Z = \{(T_i, \tilde{Y}_i)\}$ (marked PP with intensity measure $\mu \times \mathsf{P}_Y$)



Poisson construction for general Y

- Let $T_1, T_2, \ldots$ be arrival times of Poisson process with $\lambda = 1$, indep. of X
- Let $\tilde{Y}_1, \tilde{Y}_2, \ldots$ be i.i.d. $\sim \mathsf{P}_Y$ indep. of $T_1, T_2, \ldots$ and X
- Set $Z = \{(T_i, \tilde{Y}_i)\}$ (marked PP with intensity measure $\mu \times \mathsf{P}_Y$), and

$$Y = g(x, Z) = \tilde{Y}_{k(x,Z)}, \text{ where } k(x, Z) = \operatorname*{argmin}_i t_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(.|x)}(\tilde{Y}_i)$$



Example

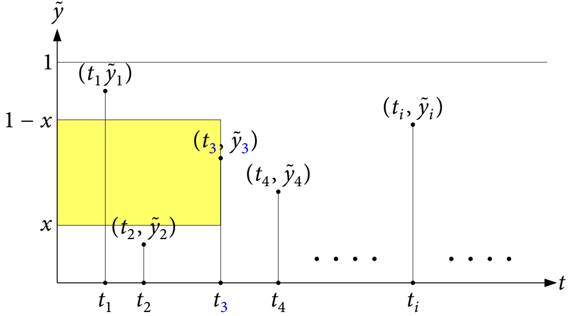
- Let $Y \sim \mathrm{Unif}[0,1]$, $Y|\{X=x\} \sim \mathrm{Unif}[x, 1-x]$, $x \in [0, 1/2]$, hence

$$\begin{aligned} k(x,z) &= \operatorname*{argmin}_i t_i \cdot \frac{f_Y(\tilde{y}_i)}{f_{Y|X}(\tilde{y}_i|x)} \text{ for } \tilde{y}_i \in [x, 1-x] \\ &= \operatorname*{argmin}_i t_i \cdot (1-2x) \text{ for } \tilde{y}_i \in [x, 1-x] \end{aligned}$$

Example

- Let $Y \sim \text{Unif}[0,1]$, $Y|\{X=x\} \sim \text{Unif}[x, 1-x]$, $x \in [0, 1/2]$, hence

$$k(x,z) = \underset{i}{\text{argmin}}\, t_i \cdot \frac{f_Y(\tilde{y}_i)}{f_{Y|X}(\tilde{y}_i|x)} \text{ for } \tilde{y}_i \in [x, 1-x]$$
$$= \underset{i}{\text{argmin}}\, t_i \cdot (1-2x) \text{ for } \tilde{y}_i \in [x, 1-x]$$
$$= \underset{i}{\text{argmin}}\, t_i \text{ for } \tilde{y}_i \in [x, 1-x],$$

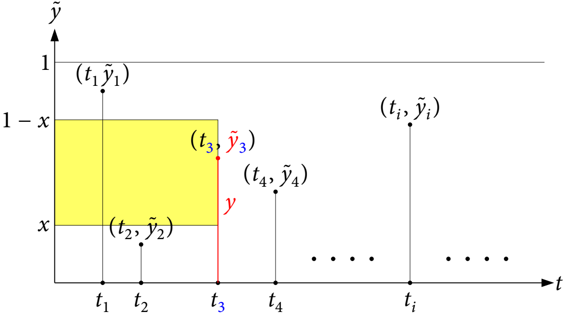


Example

- Let $Y \sim \mathrm{Unif}[0,1]$, $Y|\{X=x\} \sim \mathrm{Unif}[x, 1-x]$, $x \in [0, 1/2]$, hence

$$\begin{aligned} k(x,z) &= \operatorname*{argmin}_i t_i \cdot \frac{f_Y(\tilde{y}_i)}{f_{Y|X}(\tilde{y}_i|x)} \text{ for } \tilde{y}_i \in [x, 1-x] \\ &= \operatorname*{argmin}_i t_i \cdot (1-2x) \text{ for } \tilde{y}_i \in [x, 1-x] \\ &= \operatorname*{argmin}_i t_i \text{ for } \tilde{y}_i \in [x, 1-x], \end{aligned}$$

$$y = \tilde{y}_k$$



SFRL proof outline

Strong functional representation lemma (Li–EG 2018)

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$, and

$$I(X;Y) \le H(Y|Z) < I(X;Y) + \log(I(X;Y) + 1) + 4$$

- Poisson construction: $Z = \{(T_i, \tilde{Y}_i)\}$ marked PP with intensity measure $\mu \times \mathsf{P}_Y$,

$$Y = g(x, Z) = \tilde{Y}_{k(x,Z)}, \text{ where } k(x, Z) = \operatorname*{argmin}_i T_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}(\tilde{Y}_i)$$

SFRL proof outline

Strong functional representation lemma (Li–EG 2018)

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$, and

$$I(X; Y) \le H(Y|Z) < I(X; Y) + \log(I(X; Y) + 1) + 4$$

- Poisson construction: $Z = \{(T_i, \tilde{Y}_i)\}$ marked PP with intensity measure $\mu \times \mathsf{P}_Y$,

$$Y = g(x, Z) = \tilde{Y}_{k(x,Z)}, \text{ where } k(x, Z) = \operatorname*{argmin}_i T_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}(\tilde{Y}_i)$$

- $g(x, Z) \sim \mathsf{P}_{Y|X}(\cdot|x)$: Consider mapping $(t, y) \mapsto (t \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(y), y)$

SFRL proof outline

Strong functional representation lemma (Li–EG 2018)

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$, and

$$I(X;Y) \le H(Y|Z) < I(X;Y) + \log(I(X;Y)+1) + 4$$

- Poisson construction: $Z = \{(T_i, \tilde{Y}_i)\}$ marked PP with intensity measure $\mu \times \mathsf{P}_Y$,

$$Y = g(x,Z) = \tilde{Y}_{k(x,Z)}, \text{ where } k(x,Z) = \operatorname*{argmin}_i T_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}(\tilde{Y}_i)$$

- $g(x,Z) \sim \mathsf{P}_{Y|X}(\cdot|x)$: Consider mapping $(t, y) \mapsto (t \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(y), y)$

  By the mapping theorem, $\{(T_i \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(\tilde{Y}_i), \tilde{Y}_i)\}$ is PP with intensity measure $\mu \times \mathsf{P}_{Y|X}$

SFRL proof outline

Strong functional representation lemma (Li–EG 2018)

Given (X, Y), there exists Z independent of X and function $g(x, z)$ such that $Y = g(X, Z)$, and

$$I(X; Y) \le H(Y|Z) < I(X; Y) + \log(I(X; Y) + 1) + 4$$

- Poisson construction: $Z = \{(T_i, \tilde{Y}_i)\}$ marked PP with intensity measure $\mu \times \mathsf{P}_Y$,

$$Y = g(x, Z) = \tilde{Y}_{k(x,Z)}, \text{ where } k(x, Z) = \operatorname*{argmin}_i T_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}(\tilde{Y}_i)$$

- $g(x, Z) \sim \mathsf{P}_{Y|X}(\cdot|x)$: Consider mapping $(t, y) \mapsto (t \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(y), y)$

  By the mapping theorem, $\{(T_i \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(\tilde{Y}_i), \tilde{Y}_i)\}$ is PP with intensity measure $\mu \times \mathsf{P}_{Y|X}$

  Hence, $\Theta = \min_i T_i \cdot \frac{d\,\mathsf{P}_Y}{d\,\mathsf{P}_{Y|X}(\cdot|x)}(\tilde{Y}_i) \sim \mathrm{Exp}(1)$, $\tilde{Y}_K | \{X = x\} \sim \mathsf{P}_{Y|X}(\cdot|x)$

SFRL proof outline

- Since Y is a function of Z and K: $H(Y|Z) \leq H(K)$
- Proposition (max. $H(K)$ for fixed $\mathsf{E}(\log K)$): Let $K \in \mathbb{N}$, then

$$H(K) \leq \mathsf{E}(\log K) + \log(\mathsf{E}(\log K) + 1) + 1$$

SFRL proof outline

- Since Y is a function of Z and K: $H(Y|Z) \leq H(K)$
- Proposition (max. $H(K)$ for fixed $\mathsf{E}(\log K)$): Let $K \in \mathbb{N}$, then

$$H(K) \leq \mathsf{E}(\log K) + \log(\mathsf{E}(\log K) + 1) + 1$$

- To bound $\mathsf{E}(\log K)$, first consider $\mathsf{E}(\log K|X = x)$

  Given $\{X = x, \Theta = \theta\}$, $\{(T_i \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(\tilde{Y}_i), \tilde{Y}_i)\}_{i \neq K}$ is marked PP with intensity measure $\mu_{[\theta,\infty)} \times \mathsf{P}_{Y|X}(.|x)$

SFRL proof outline

- Since Y is a function of Z and K: $H(Y|Z) \le H(K)$
- Proposition (max. $H(K)$ for fixed $\mathsf{E}(\log K)$): Let $K \in \mathbb{N}$, then

$$H(K) \le \mathsf{E}(\log K) + \log(\mathsf{E}(\log K) + 1) + 1$$

- To bound $\mathsf{E}(\log K)$, first consider $\mathsf{E}(\log K | X = x)$

  Given $\{X = x, \Theta = \theta\}$, $\{(T_i \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(\tilde{Y}_i), \tilde{Y}_i)\}_{i \neq K}$ is marked PP with intensity measure $\mu_{[\theta,\infty)} \times \mathsf{P}_{Y|X}(.|x)$

  By mapping theorem for $(t, y) \mapsto (t \cdot d\,\mathsf{P}_{Y|X} / d\,\mathsf{P}_Y(\cdot|x)(y), y)$, $\{(T_i, \tilde{Y}_i)\}_{i \neq K}$ is marked PP

SFRL proof outline

- Since Y is a function of Z and K: $H(Y|Z) \leq H(K)$
- Proposition (max. $H(K)$ for fixed $\mathsf{E}(\log K)$): Let $K \in \mathbb{N}$, then

$$H(K) \leq \mathsf{E}(\log K) + \log(\mathsf{E}(\log K) + 1) + 1$$

- To bound $\mathsf{E}(\log K)$, first consider $\mathsf{E}(\log K|X = x)$

  Given $\{X = x, \Theta = \theta\}$, $\{(T_i \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(\tilde{Y}_i), \tilde{Y}_i)\}_{i \neq K}$ is marked PP with intensity measure $\mu_{[\theta,\infty)} \times \mathsf{P}_{Y|X}(.|x)$

  By mapping theorem for $(t, y) \mapsto (t \cdot d\,\mathsf{P}_{Y|X} / d\,\mathsf{P}_Y(\cdot|x)(y), y)$, $\{(T_i, \tilde{Y}_i)\}_{i \neq K}$ is marked PP

  Hence, given $\{X = x, \Theta = \theta, \tilde{Y}_K = \tilde{y}\}$, $K - 1 \sim \text{Poisson}(\lambda(\tilde{y}))$

SFRL proof outline

- Since Y is a function of Z and K: $H(Y|Z) \le H(K)$
- Proposition (max. $H(K)$ for fixed $\mathsf{E}(\log K)$): Let $K \in \mathbb{N}$, then

$$H(K) \le \mathsf{E}(\log K) + \log(\mathsf{E}(\log K) + 1) + 1$$

- To bound $\mathsf{E}(\log K)$, first consider $\mathsf{E}(\log K|X = x)$

  Given $\{X = x, \Theta = \theta\}$, $\{(T_i \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(\tilde{Y}_i), \tilde{Y}_i)\}_{i \neq K}$ is marked PP with intensity measure $\mu_{[\theta,\infty)} \times \mathsf{P}_{Y|X}(.|x)$

  By mapping theorem for $(t, y) \mapsto (t \cdot d\,\mathsf{P}_{Y|X} / d\,\mathsf{P}_Y(\cdot|x)(y), y)$, $\{(T_i, \tilde{Y}_i)\}_{i \neq K}$ is marked PP

  Hence, given $\{X = x, \Theta = \theta, \tilde{Y}_K = \tilde{y}\}$, $K - 1 \sim \mathrm{Poisson}(\lambda(\tilde{y}))$

  It's not difficult to show: $\lambda(\tilde{y}) \le \theta \cdot d\,\mathsf{P}_{Y|X}(\cdot|x)/d\,\mathsf{P}_Y(\tilde{y})$

SFRL proof outline

- Since Y is a function of Z and K: $H(Y|Z) \leq H(K)$
- Proposition (max. $H(K)$ for fixed $\mathsf{E}(\log K)$): Let $K \in \mathbb{N}$, then

$$H(K) \leq \mathsf{E}(\log K) + \log(\mathsf{E}(\log K) + 1) + 1$$

- To bound $\mathsf{E}(\log K)$, first consider $\mathsf{E}(\log K|X = x)$

  Given $\{X = x, \Theta = \theta\}$, $\{(T_i \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(\tilde{Y}_i), \tilde{Y}_i)\}_{i \neq K}$ is marked PP with intensity measure $\mu_{[\theta,\infty)} \times \mathsf{P}_{Y|X}(.|x)$

  By mapping theorem for $(t, y) \mapsto (t \cdot d\,\mathsf{P}_{Y|X} / d\,\mathsf{P}_Y(\cdot|x)(y), y)$, $\{(T_i, \tilde{Y}_i)\}_{i \neq K}$ is marked PP

  Hence, given $\{X = x, \Theta = \theta, \tilde{Y}_K = \tilde{y}\}$, $K - 1 \sim \mathrm{Poisson}(\lambda(\tilde{y}))$

  It's not difficult to show: $\lambda(\tilde{y}) \leq \theta \cdot d\,\mathsf{P}_{Y|X}(\cdot|x)/d\,\mathsf{P}_Y(\tilde{y})$

  We can now show: $\mathsf{E}(\log K\,|\,X = x) \leq D(\mathsf{P}_{Y|X}(\cdot|x)\,||\,\mathsf{P}_Y) + e^{-1}\log e + 1$

SFRL proof outline

- Since Y is a function of Z and K: $H(Y|Z) \leq H(K)$
- Proposition (max. $H(K)$ for fixed $\mathsf{E}(\log K)$): Let $K \in \mathbb{N}$, then

$$H(K) \leq \mathsf{E}(\log K) + \log(\mathsf{E}(\log K) + 1) + 1$$

- To bound $\mathsf{E}(\log K)$, first consider $\mathsf{E}(\log K|X = x)$

  Given $\{X = x, \Theta = \theta\}$, $\{(T_i \cdot d\,\mathsf{P}_Y / d\,\mathsf{P}_{Y|X}(\cdot|x)(\tilde{Y}_i), \tilde{Y}_i)\}_{i \neq K}$ is marked PP with intensity measure $\mu_{[\theta,\infty)} \times \mathsf{P}_{Y|X}(.|x)$

  By mapping theorem for $(t, y) \mapsto (t \cdot d\,\mathsf{P}_{Y|X} / d\,\mathsf{P}_Y(\cdot|x)(y), y)$, $\{(T_i, \tilde{Y}_i)\}_{i \neq K}$ is marked PP

  Hence, given $\{X = x, \Theta = \theta, \tilde{Y}_K = \tilde{y}\}$, $K - 1 \sim \mathrm{Poisson}(\lambda(\tilde{y}))$

  It's not difficult to show: $\lambda(\tilde{y}) \leq \theta \cdot d\,\mathsf{P}_{Y|X}(\cdot|x)/d\,\mathsf{P}_Y(\tilde{y})$

  We can now show: $\mathsf{E}(\log K|X = x) \leq D(\mathsf{P}_{Y|X}(\cdot|x)||\,\mathsf{P}_Y) + e^{-1}\log e + 1$

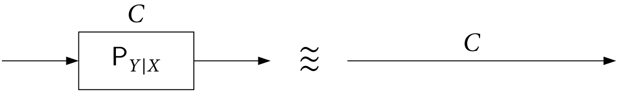
- Taking expect. over X and substituting into Proposition complete proof

Applications of SFRL

- Upper bound on rate of one-shot (exact) channel simulation
- One-shot lossy compression
- Minimax learning for distributed inference (Li–Wu–Özgür–EG 2018)

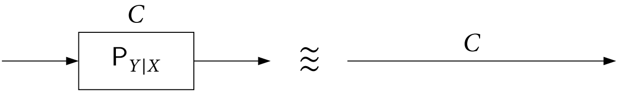
Background on channel simulation

- Shannon (1948) channel capacity theorem can be interpreted as:
  DMC with capacity C can simulate noiseless channel with capacity C

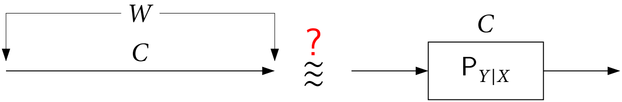


Background on channel simulation

- Shannon (1948) channel capacity theorem can be interpreted as:
  DMC with capacity C can simulate noiseless channel with capacity C

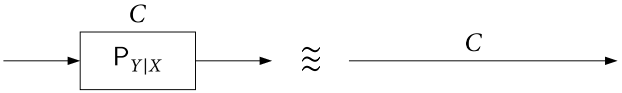


- Bennett–Shor–Smolin–Thapliyal (2002) asked the reverse question:
  Can noiseless channel with capacity C and common randomness simulate any DMC with capacity C?

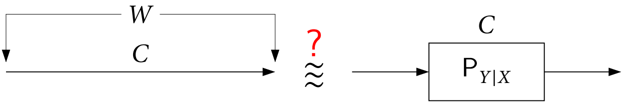


Background on channel simulation

- Shannon (1948) channel capacity theorem can be interpreted as:
  DMC with capacity C can simulate noiseless channel with capacity C

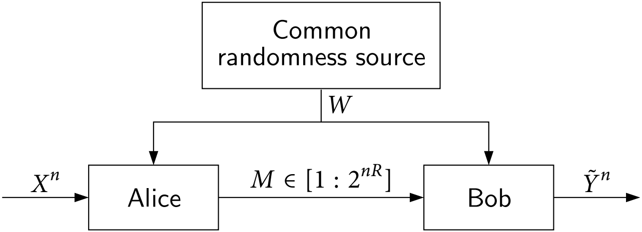


- Bennett–Shor–Smolin–Thapliyal (2002) asked the reverse question:
  Can noiseless channel with capacity C and common randomness simulate any DMC with capacity C?



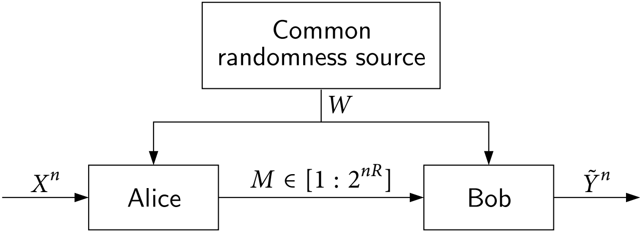
- Their motivation was to answer this question for entanglement-assisted quantum channels

Approximate channel simulation



- W unlimited common randomness; X arbitrary process; $p(y|x)$ DMC
- Alice maps every (x^n, w) pair into an index $m(x^n, w) \in [1 : 2^{nR}]$
- Bob generates $\tilde{Y}^n(m(x^n, W), W) \sim q(y^n|x^n)$

Approximate channel simulation



- W unlimited common randomness; X arbitrary process; $p(y|x)$ DMC
- Alice maps every (x^n, w) pair into an index $m(x^n, w) \in [1:2^{nR}]$
- Bob generates $\tilde{Y}^n(m(x^n, W), W) \sim q(y^n|x^n)$
- R is achievable if there exists sequence of simulation schemes such that

$$\lim_{n\to\infty} \left\| p(x^n)q(y^n|x^n) - p(x^n)\prod_{i=1}^{n} p_{Y|X}(y_i|x_i) \right\|_{\mathrm{TV}} = 0$$

- Optimal (approx.) simulation rate $R^*_{\mathrm{ch-sim}}$ is inf. over achievable rates

Approximate channel simulation

Theorem (Bennett–Shor–Smolin–Thapliyal 2002)

$$R^*_{\text{ch-sim}} = \max_{p(x)} I(X;Y) \quad \text{(capacity of DMC)}$$

- Hence reverse Shannon channel capacity theorem holds for DMC

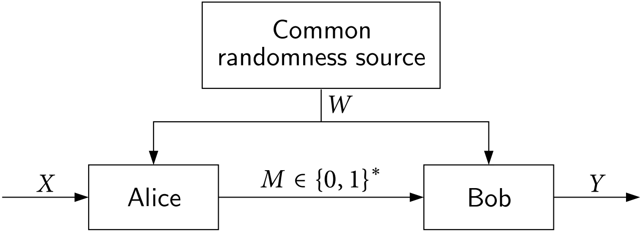
Approximate channel simulation

Theorem (Bennett–Shor–Smolin–Thapliyal 2002)

$$R^*_{\mathrm{ch-sim}} = \max_{p(x)} I(X;Y) \quad \text{(capacity of DMC)}$$

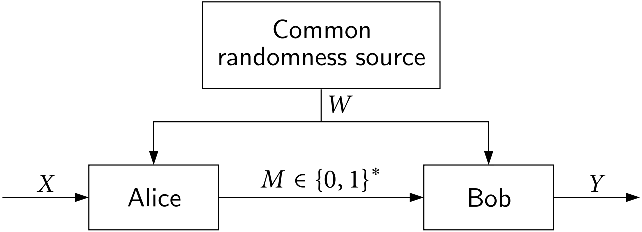
- Hence reverse Shannon channel capacity theorem holds for DMC
- They also established partial results for certain quantum channels
- Follow on work (Cuff 2013, Bennett–Devetak–Harrow–Shor–Winter 2014)

One-shot exact channel simulation



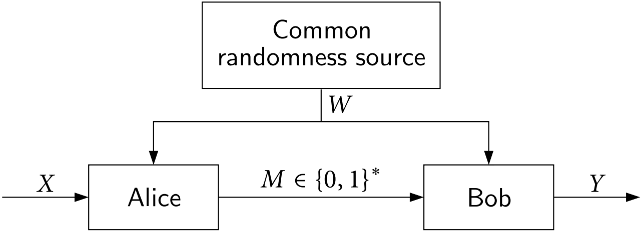
- W unlimited common randomness; $X \sim \mathsf{P}_X$, $\mathsf{P}_{Y|X}$ given channel

One-shot exact channel simulation



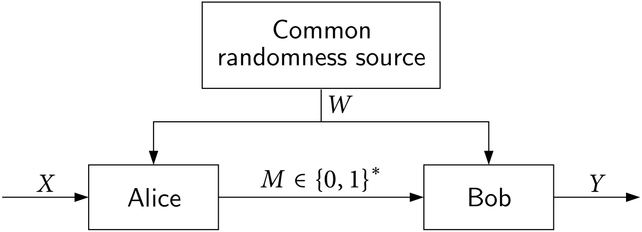
- W unlimited common randomness; $X \sim \mathsf{P}_X$, $\mathsf{P}_{Y|X}$ given channel
- For each w, Alice uses prefix code to map each x into $m(x, w) \in \{0,1\}^*$
- Bob generates $Y = y(m(x, W), W) \sim \mathsf{P}_{Y|X}(.|x)$
- Let L be the length of the index M

One-shot exact channel simulation



- W unlimited common randomness; $X \sim \mathsf{P}_X$, $\mathsf{P}_{Y|X}$ given channel
- For each w, Alice uses prefix code to map each x into $m(x,w) \in \{0,1\}^*$
- Bob generates $Y = y(m(x,W),W) \sim \mathsf{P}_{Y|X}(.|x)$
- Let L be the length of the index M
- Optimal average simulation rate is $\bar{R}^*_{\text{ch-cim}} = \inf_{\text{generators}} \mathsf{E}(L)$

One-shot exact channel simulation



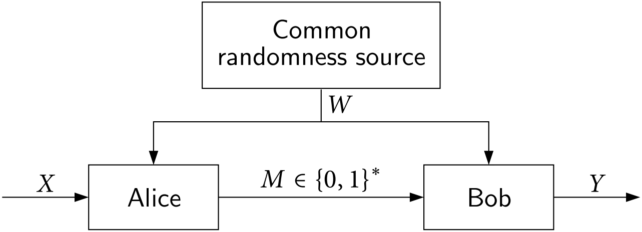
Theorem (Harsha–Jain–McAllester–Radhakrishnan 2010)

For discrete $X \sim p(x)$, and DMC $p(y|x)$,

$$I(X;Y) \le \bar{R}^*_{\text{ch}-\text{sim}} \le I(X;Y) + (1+\epsilon)\log(I(X;Y)+1) + c_\epsilon$$

- More generally they showed for any x: $\bar{R}^*_{\text{ch}-\text{sim}} \le C + (1+\epsilon)\log(C+1) + c_\epsilon$
- Proof uses rejection sampling and is quite involved

One-shot exact channel simulation



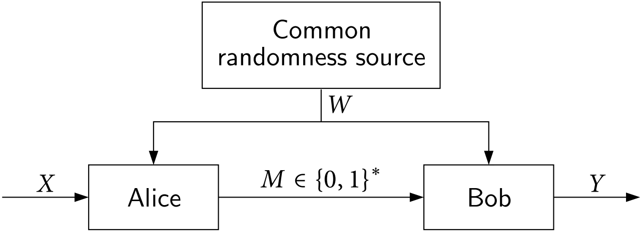
Theorem (Li–EG 2018)

For $X \sim \mathrm{P}_X$, and general memoryless channel $\mathrm{P}_{Y|X}$,

$$I(X;Y) \le \bar{R}^*_{\mathrm{ch-sim}} < I(X;Y) + \log(I(X;Y)+1) + 5$$

- Proof of upper bound uses SFRL
- Can be extended to arbitrary x case

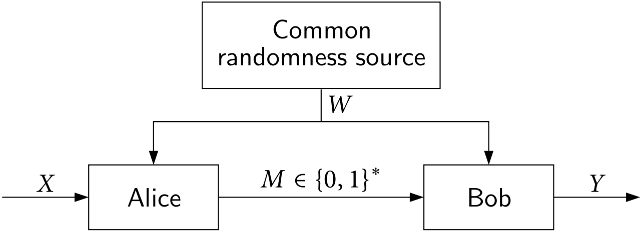
Proof of upper bound using SFRL



- By SFRL, there exists W indep. of X such that $Y = g(X, W)$, and

$$H(Y|W) < I(X;Y) + \log(I(X;Y) + 1) + 4$$

Proof of upper bound using SFRL

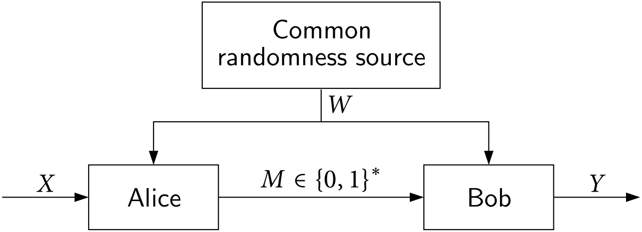


- By SFRL, there exists W indep. of X such that $Y = g(X, W)$, and

$$H(Y|W) < I(X;Y) + \log(I(X;Y) + 1) + 4$$

- For each w, map $y(x, w)$ into $m(x, w)$ using Huffman code for $p_{Y|W}(y|w)$
- Codes $\{y(x, w), m(x, w)\}$ are provided to Alice and Bob

Proof of upper bound using SFRL

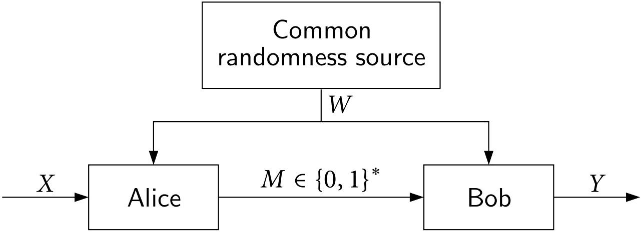


- By SFRL, there exists W indep. of X such that $Y = g(X, W)$, and

$$H(Y|W) < I(X;Y) + \log(I(X;Y) + 1) + 4$$

- For each w, map $y(x, w)$ into $m(x, w)$ using Huffman code for $p_{Y|W}(y|w)$
- Codes $\{y(x, w), m(x, w)\}$ are provided to Alice and Bob
- Given (x, w), Alice computes $y(x, w)$; given (m, w), Bob recovers y

Proof of upper bound using SFRL

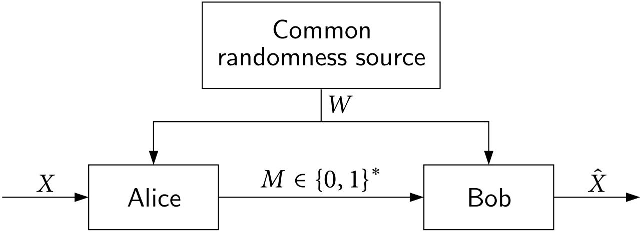


- By SFRL, there exists W indep. of X such that $Y = g(X, W)$, and

$$H(Y|W) < I(X;Y) + \log(I(X;Y)+1) + 4$$

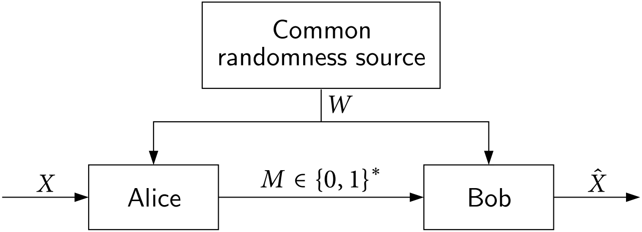
- For each w, map $y(x,w)$ into $m(x,w)$ using Huffman code for $p_{Y|W}(y|w)$
- Codes $\{y(x,w), m(x,w)\}$ are provided to Alice and Bob
- Given (x,w), Alice computes $y(x,w)$; given (m,w), Bob recovers y
- Hence, $\bar{R}^*_{\text{ch-sim}} \le \mathsf{E}(L) < H(Y|W) + 1 < I(X;Y) + \log(I(X;Y)+1) + 5$

One-shot lossy source coding



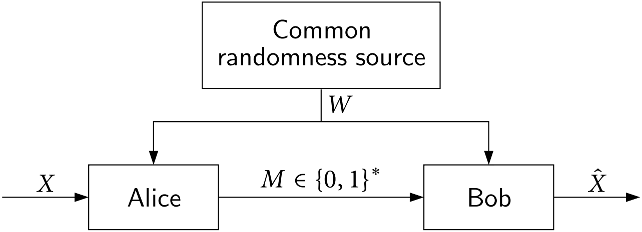
- $X \sim \mathsf{P}_X$, $\hat{\mathcal{X}}$ reproduction alphabet, $d(x, \hat{x}) \geq 0$ distortion measure

One-shot lossy source coding



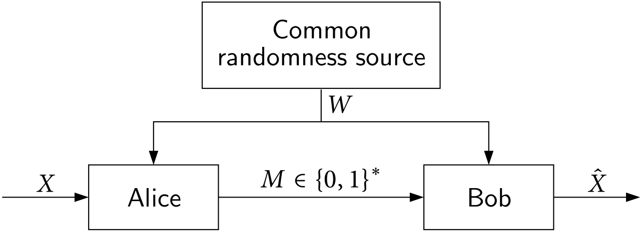
- $X \sim \mathsf{P}_X$, $\hat{\mathcal{X}}$ reproduction alphabet, $d(x, \hat{x}) \geq 0$ distortion measure
- Alice maps each (x, w) into $m(x, w)$; let L be length of M
- Bob maps each (m, w) into estimate $\hat{x}(m, w)$

One-shot lossy source coding



- $X \sim \mathsf{P}_X$, $\hat{\mathcal{X}}$ reproduction alphabet, $d(x, \hat{x}) \geq 0$ distortion measure
- Alice maps each (x, w) into $m(x, w)$; let L be length of M
- Bob maps each (m, w) into estimate $\hat{x}(m, w)$
- $(\bar{R}, D)$ is achievable if there exits code with $\bar{R} = \mathsf{E}(L)$, $\mathsf{E}(d(X, \hat{X})) \leq D$
- Avg rate-dist. function $\bar{R}(D)$ is inf over all achievable $\bar{R}$: $\mathsf{E}(d(X, \hat{X})) \leq D$

One-shot lossy source coding



- Avg rate-dist. function $\bar{R}(D)$ is inf over all achievable $\bar{R}$: $\mathsf{E}(d(X,\hat{X})) \le D$

Theorem (Li–EG 2018)

$$R(D) \le \bar{R}(D) < R(D) + \log(R(D)+1) + 5,$$

where $R(D) = \inf_{P_{\hat{X}|X}:\, \mathsf{E}(d(X,\hat{X})) \le D} I(X;\hat{X})$ (rate-dist. function for asymptotic case)

Proof of upper bound using SFRL

- Let $\hat{X}$ attain $I(X;\hat{X}) = R(D)$ and $\mathsf{E}(d(X,\hat{X})) \le D$

Proof of upper bound using SFRL

- Let $\hat{X}$ attain $I(X;\hat{X}) = R(D)$ and $\mathsf{E}(d(X,\hat{X})) \le D$
- By SFRL, there exists W indep. of X s.t. $\hat{X} = g(X,W)$ and

$$H(\hat{X}|W) < I(X;\hat{X}) + \log(I(X;\hat{X}) + 1) + 4 = R(D) + \log(R(D) + 1) + 4 \ (*)$$

- By Carathéadory's theorem, $|\mathcal{W}| \le 2$ suffices to satisfy $(*)$, $\mathsf{E}(d(X,\hat{X})) \le D$

Proof of upper bound using SFRL

- Let $\hat{X}$ attain $I(X;\hat{X}) = R(D)$ and $\mathsf{E}(d(X,\hat{X})) \le D$
- By SFRL, there exists W indep. of X s.t. $\hat{X} = g(X,W)$ and

$$H(\hat{X}|W) < I(X;\hat{X}) + \log(I(X;\hat{X}) + 1) + 4 = R(D) + \log(R(D) + 1) + 4 \ (*)$$

- By Carathéadory's theorem, $|\mathcal{W}| \le 2$ suffices to satisfy $(*)$, $\mathsf{E}(d(X,\hat{X})) \le D$
- For each $w \in \{1,2\}$, use Huffman code to map $\hat{x}$ into $m(\hat{x},w) \in \{0,1\}^*$
- Codes $\{\hat{x}(x,w), m(\hat{x},w)\}$ are provided to Alice and Bob

Proof of upper bound using SFRL

- Let $\hat{X}$ attain $I(X;\hat{X}) = R(D)$ and $\mathsf{E}(d(X,\hat{X})) \le D$
- By SFRL, there exists W indep. of X s.t. $\hat{X} = g(X,W)$ and

$$H(\hat{X}|W) < I(X;\hat{X}) + \log(I(X;\hat{X})+1) + 4 = R(D) + \log(R(D)+1) + 4 \ (*)$$

- By Carathéodory's theorem, $|\mathcal{W}| \le 2$ suffices to satisfy $(*)$, $\mathsf{E}(d(X,\hat{X})) \le D$
- For each $w \in \{1,2\}$, use Huffman code to map $\hat{x}$ into $m(\hat{x},w) \in \{0,1\}^*$
- Codes $\{\hat{x}(x,w), m(\hat{x},w)\}$ are provided to Alice and Bob
- Alice maps (x,w) into m
- Bob recovers $\hat{x}(x,w)$

Proof of upper bound using SFRL

- Let $\hat{X}$ attain $I(X;\hat{X}) = R(D)$ and $\mathsf{E}(d(X,\hat{X})) \le D$
- By SFRL, there exists W indep. of X s.t. $\hat{X} = g(X,W)$ and

$$H(\hat{X}|W) < I(X;\hat{X}) + \log(I(X;\hat{X}) + 1) + 4 = R(D) + \log(R(D) + 1) + 4 \ (*)$$

- By Carathéadory's theorem, $|\mathcal{W}| \le 2$ suffices to satisfy $(*)$, $\mathsf{E}(d(X,\hat{X})) \le D$
- For each $w \in \{1,2\}$, use Huffman code to map $\hat{x}$ into $m(\hat{x},w) \in \{0,1\}^*$
- Codes $\{\hat{x}(x,w), m(\hat{x},w)\}$ are provided to Alice and Bob
- Alice maps (x,w) into m
- Bob recovers $\hat{x}(x,w)$
- Hence, $\bar{R}(D) \le \mathsf{E}(L) < H(\hat{X}|W) + 1 < R(D) + \log(R(D)+1) + 5,\ \mathsf{E}(d(X,\hat{X})) \le D$

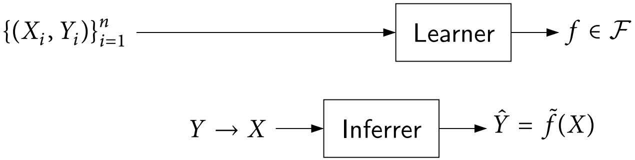
Related work

- Pinkston (1967) studied variable-length finite blocklength lossy compression for i.i.d. source, per-letter distortion
- Zhang–Yang–Wei (1997) established similar order bound to ours for finite blocklength
- Kostina–Polyanskiy–Verdú (2015) studied variable length finite blocklength lossy compression with prob. of distortion constraint
- Our coding scheme resembles Song–Cuff–Poor (2016) likelihood encoder

Applications of SFRL

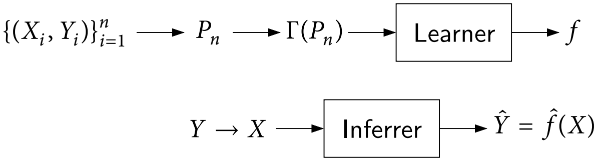
- Upper bound on rate of one-shot (exact) channel simulation
- One-shot lossy compression
- Minimax learning for distributed inference (Li–Wu–Özgür–EG 2018)

Supervised learning



- Risk function: $l(y, \hat{y})$, P_n empirical pmf of (X, Y), function class $\mathcal{F}$
- Empirical risk minimization: choose $\tilde{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \mathsf{E}_{P_n}(l(Y, \hat{Y}))$

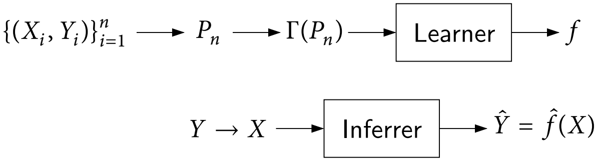
Minimax learning



- Minimax learning: choose $\hat{f} = \operatorname{argmin}_f \max_{P \in \Gamma(P_n)} \mathsf{E}_P(l(Y, \hat{Y}))$

  $\Gamma(P_n)$: ambiguity set around P_n, e.g.,
  - Set of pmfs with same 1st, 2nd moments as P_n (Farnia–Tse 2016)
  - f-divergence, Wasserstein ball (Namkoong–Duchi 2017, Lee–Raginsky 2017)

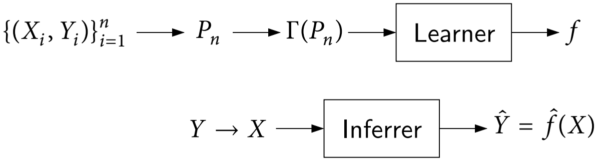
Minimax learning



- Minimax learning: choose $\hat{f} = \operatorname{argmin}_f \max_{P \in \Gamma(P_n)} \mathsf{E}_P(l(Y, \hat{Y}))$

  $\Gamma(P_n)$: ambiguity set around P_n, e.g.,
  - Set of pmfs with same 1st, 2nd moments as P_n (Farnia–Tse 2016)
  - f-divergence, Wasserstein ball (Namkoong–Duchi 2017, Lee–Raginsky 2017)
- If X, Y discrete, Γ convex, closed (Farnia–Tse 2016):
  - First find $p^* = \operatorname{argmax}_{p \in \Gamma(p_n)} \min_f \mathsf{E}_p(l(Y, \hat{Y}))$, use it to find $\hat{f} = \operatorname{argmin}_f \mathsf{E}_{p^*}(l(Y, \hat{Y}))$

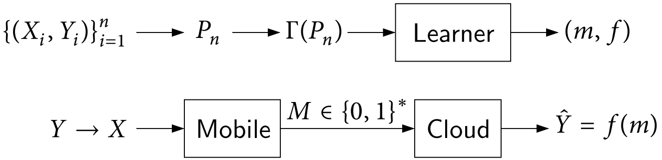
Minimax learning



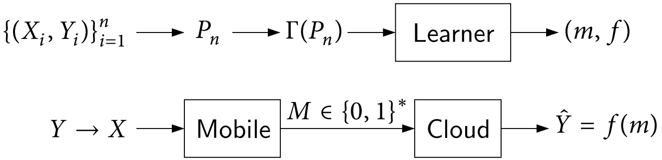
- Minimax learning: choose $\hat{f} = \underset{f}{\operatorname{argmin}} \max_{P \in \Gamma(P_n)} \mathsf{E}_P(l(Y, \hat{Y}))$

  $\Gamma(P_n)$: ambiguity set around P_n, e.g.,
  - Set of pmfs with same 1st, 2nd moments as P_n (Farnia–Tse 2016)
  - f-divergence, Wasserstein ball (Namkoong–Duchi 2017, Lee–Raginsky 2017)
- If X, Y discrete, Γ convex, closed (Farnia–Tse 2016):
  - First find $p^* = \underset{p \in \Gamma(p_n)}{\operatorname{argmax}} \min_f \mathsf{E}_p(l(Y, \hat{Y}))$, use it to find $\hat{f} = \underset{f}{\operatorname{argmin}} \mathsf{E}_{p^*}(l(Y, \hat{Y}))$
  - Recovers linear/logistic regression for suitable l, Γ

Minimax learning for distributed inference

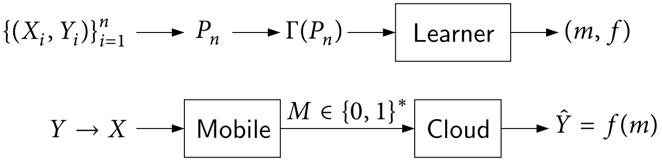


Minimax learning for distributed inference

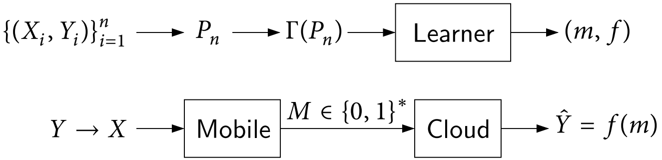


- Assume common randomness W available between cloud/mobile
- Mobile maps every (x, w) into index $m(x, w)$
- Cloud maps (m, w) into an estimate $\hat{y} = f(m, w)$
- Let T be the length of M

Minimax learning for distributed inference



- Assume common randomness W available between cloud/mobile
- Mobile maps every (x, w) into index $m(x, w)$
- Cloud maps (m, w) into an estimate $\hat{y} = f(m, w)$
- Let T be the length of M
- Minimax risk-rate cost: $L_\lambda^* = \inf_{m,f} \sup_{P \in \Gamma} \left[\mathsf{E}_P(l(Y, \hat{Y})) + \lambda \, \mathsf{E}_P(T) \right]$



- Let T be the length of M
- Minimax risk-rate cost: $L_\lambda^* = \inf_{m,f} \sup_{P\in\Gamma} \left[\mathsf{E}_P(l(Y,\hat{Y})) + \lambda\, \mathsf{E}_P(T) \right]$

Theorem (Li–Wu–Özgür–EG 2018)

Let Γ be convex, then

$$L_\lambda^* \geq \inf_{\hat{P}_{\hat{Y}|X}} \sup_{P\in\Gamma} \left[\mathsf{E}_P(l(Y,\hat{Y})) + \lambda I(X;\hat{Y}) \right]$$

$$L_\lambda^* < \inf_{\hat{P}_{\hat{Y}|X}} \sup_{P\in\Gamma} \left[\mathsf{E}_P(l(Y,\hat{Y})) + \lambda (I(X;\hat{Y}) + 2\log(I(X;\hat{Y})+1) + 6) \right]$$

Proof outline of upper bound

Theorem (Li–Wu–Özgür–EG 2018)

Let Γ be convex, then

$$L_\lambda^* < \inf_{\hat{P}_{\hat{Y}|X}} \sup_{P\in\Gamma} \left[\mathsf{E}_P(l(Y,\hat{Y})) + \lambda(I(X;\hat{Y}) + 2\log(I(X;\hat{Y})+1) + 6)\right]$$

- For $\Gamma = \{P\}$, problem reduces to one-shot noisy lossy compression

  Proof essentially same as for one-shot lossy compression via SFRL,

$$L_\lambda^* < \inf_{\hat{P}_{\hat{Y}|X}} \left[\mathsf{E}(l(Y,\hat{Y})) + \lambda(I(X;\hat{Y}) + \log(I(X;\hat{Y})+1) + 5)\right]$$

Proof outline of upper bound

Theorem (Li–Wu–Özgür–EG 2018)

Let Γ be convex, then

$$L_\lambda^* < \inf_{\hat{P}_{\hat{Y}|X}} \sup_{P\in\Gamma} \left[\mathsf{E}_P(l(Y,\hat{Y})) + \lambda(I(X;\hat{Y}) + 2\log(I(X;\hat{Y})+1) + 6)\right]$$

- For general Γ, we need refined version of SFRL:

  For $\mathsf{P}_{\hat{Y}|X}$, $\tilde{\mathsf{P}}_{\hat{Y}}$, there exists r.v. W, two functions $k(x,w)\in\mathbb{N}$, $\hat{y}(k,w)$:

$$\hat{y}(k(x,W),W) \sim \mathsf{P}_{\hat{Y}|X}$$
$$\mathsf{E}(\log k(x,W)) \le D(\mathsf{P}_{\hat{Y}|X}(.|x)||\tilde{\mathsf{P}}_{\hat{Y}}) + 1.6$$

- Encode K using Elias (1975) codes: $\mathsf{E}(T) \le \mathsf{E}(\log K) + 2\log(\mathsf{E}(\log K)+1)+1$
- Rest of proof is technical, see details in (Li–Wu–Özgür–EG 2018)

Principle of max risk-information cost

- Minimax risk-rate cost: $L_\lambda^* = \inf_{f,m} \sup_{P \in \Gamma} \big(\mathsf{E}_p(l(Y, \hat{Y})) + \lambda\, \mathsf{E}_p(T) \big)$

Principle of max risk-information cost

- Minimax risk-rate cost: $L_\lambda^* = \inf_{f,m} \sup_{P \in \Gamma} \big(\mathsf{E}_p(l(Y, \hat{Y})) + \lambda \, \mathsf{E}_p(T) \big)$
- Minimax risk-information cost: $\bar{L}_\lambda^* = \inf_{P_{\hat{Y}|X}} \sup_{P \in \Gamma} \big(\mathsf{E}(l(Y, \hat{Y})) + \lambda I(X; \hat{Y}) \big)$

Principle of max risk-information cost

- Minimax risk-rate cost: $L_\lambda^* = \inf_{f,m} \sup_{P\in\Gamma} \big(\mathsf{E}_p(l(Y,\hat{Y})) + \lambda\, \mathsf{E}_p(T) \big)$
- Minimax risk-information cost: $\bar{L}_\lambda^* = \inf_{P_{\hat{Y}|X}} \sup_{P\in\Gamma} \big(\mathsf{E}(l(Y,\hat{Y})) + \lambda I(X;\hat{Y}) \big)$
- If $X, Y, \hat{Y}$ are finite, Γ convex and closed, by Sion's theorem:

$$\bar{L}_\lambda^* = \max_{p\in\Gamma} \min_{p_{\hat{Y}|X}} \big(\mathsf{E}(l(Y,\hat{Y})) + \lambda I(X;\hat{Y}) \big)$$

Principle of max risk-information cost

- Minimax risk-rate cost: $L_\lambda^* = \inf_{f,m} \sup_{P\in\Gamma} \big(\mathsf{E}_p(l(Y,\hat{Y})) + \lambda\, \mathsf{E}_p(T) \big)$
- Minimax risk-information cost: $\bar{L}_\lambda^* = \inf_{p_{\hat{Y}|X}} \sup_{P\in\Gamma} \big(\mathsf{E}(l(Y,\hat{Y})) + \lambda I(X;\hat{Y}) \big)$
- If $X, Y, \hat{Y}$ are finite, Γ convex and closed, by Sion's theorem:

$$\bar{L}_\lambda^* = \max_{p\in\Gamma} \min_{p_{\hat{Y}|X}} \big(\mathsf{E}(l(Y,\hat{Y})) + \lambda I(X;\hat{Y}) \big)$$

- To design robust descriptor-estimator pair that works for every $p \in \Gamma$,
  - First find: $p^* = \operatorname{argmax}_{p\in\Gamma} \min_{p_{\hat{Y}|X}} \big(\mathsf{E}(l(Y,\hat{Y})) + \lambda I(X;\hat{Y}) \big)$
  - Then find: $p^*_{\hat{Y}|X} = \operatorname{argmin}_{p_{\hat{Y}|X}} \big(\mathsf{E}_{p^*}(l(Y,\hat{Y}) + \lambda I_{p^*}(X;\hat{Y}) \big)$
- Extends maximum conditional entropy principle in (Farnia–Tse 2016)

Linear regression

- Let $\mathbf{X} \in \mathbb{R}^d$, $Y, \hat{Y} \in \mathbb{R}$, $l(y, \hat{y}) = (y - \hat{y})^2$, $\mathsf{E}(\mathbf{X}) = \mathbf{0}$, $\mathsf{E}(Y) = 0$

$$\Gamma = \{P_{\mathbf{X},Y} : \ \mathsf{E}(\mathbf{X}) = \mathbf{0}, \ \mathsf{E}(Y) = 0, \ \Sigma_{\mathbf{X}}, \ C_{\mathbf{X}Y}, \ \text{same as } P_n\}$$

- Let $\mathbf{X} \in \mathbb{R}^d$, $Y, \hat{Y} \in \mathbb{R}$, $l(y, \hat{y}) = (y - \hat{y})^2$, $\mathsf{E}(\mathbf{X}) = \mathbf{0}$, $\mathsf{E}(Y) = 0$

$$\Gamma = \{P_{\mathbf{X},Y} : \mathsf{E}(\mathbf{X}) = \mathbf{0},\ \mathsf{E}(Y) = 0,\ \Sigma_{\mathbf{X}},\ C_{\mathbf{X}Y},\ \text{same as } P_n\}$$

- Minimax solution: $P^*_{\mathbf{X},Y}$ Gaussian with same mean, covariance as P_n,

$$\hat{Y} = \begin{cases} a \cdot C^t_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}\mathbf{X} + Z & \text{if } a > 0, \\ 0 & \text{otherwise} \end{cases}$$

$$\tilde{L}^*_\lambda = \begin{cases} \sigma^2_Y - C^t_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y} - \frac{\lambda}{2}\log e(1-a) & \text{if } a > 0, \\ \sigma^2_Y & \text{otherwise,} \end{cases}$$

$$a = 1 - \frac{\lambda \log e}{2C^t_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y}}, \quad Z \sim \mathrm{N}\big(0, a\lambda \log e/2\big) \text{ independent of } \mathbf{X}$$

Linear regression

- Let $\mathbf{X} \in \mathbb{R}^d$, $Y, \hat{Y} \in \mathbb{R}$, $l(y, \hat{y}) = (y - \hat{y})^2$, $\mathsf{E}(\mathbf{X}) = \mathbf{0}$, $\mathsf{E}(Y) = 0$

$$\Gamma = \{P_{\mathbf{X},Y} : \mathsf{E}(\mathbf{X}) = \mathbf{0},\ \mathsf{E}(Y) = 0,\ \Sigma_{\mathbf{X}},\ C_{\mathbf{X}Y},\ \text{same as } P_n\}$$

- Minimax solution: $P^*_{\mathbf{X},Y}$ Gaussian with same mean, covariance as P_n,

$$\hat{Y} = \begin{cases} a \cdot C^t_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}\mathbf{X} + Z & \text{if } a > 0, \\ 0 & \text{otherwise} \end{cases}$$

$$\tilde{L}^*_\lambda = \begin{cases} \sigma^2_Y - C^t_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y} - \frac{\lambda}{2}\log e(1-a) & \text{if } a > 0, \\ \sigma^2_Y & \text{otherwise,} \end{cases}$$

$$a = 1 - \frac{\lambda \log e}{2C^t_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y}}, \quad Z \sim \mathrm{N}\big(0, a\lambda\log e/2\big) \text{ independent of } \mathbf{X}$$

- $\lambda = 0$ (no rate constraint) $\Rightarrow$ linear regression (Farnia–Tse 2016)

Linear regression

- Let $\mathbf{X} \in \mathbb{R}^d$, $Y, \hat{Y} \in \mathbb{R}$, $l(y, \hat{y}) = (y - \hat{y})^2$, $\mathsf{E}(\mathbf{X}) = \mathbf{0}$, $\mathsf{E}(Y) = 0$

$$\Gamma = \{P_{\mathbf{X},Y} : \mathsf{E}(\mathbf{X}) = \mathbf{0},\ \mathsf{E}(Y) = 0,\ \Sigma_{\mathbf{X}},\ C_{\mathbf{X}Y},\ \text{same as } P_n\}$$

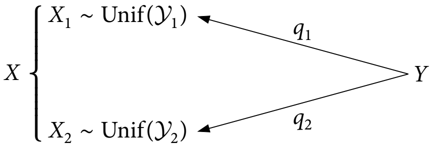
- Minimax solution: $P^*_{\mathbf{X},Y}$ Gaussian with same mean, covariance as P_n,

$$\hat{Y} = \begin{cases} a \cdot C^t_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}\mathbf{X} + Z & \text{if } a > 0, \\ 0 & \text{otherwise} \end{cases}$$

$$\tilde{L}^*_\lambda = \begin{cases} \sigma^2_Y - C^t_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y} - \frac{\lambda}{2}\log e(1-a) & \text{if } a > 0, \\ \sigma^2_Y & \text{otherwise,} \end{cases}$$

$$a = 1 - \frac{\lambda \log e}{2C^t_{\mathbf{X}Y}\Sigma^{-1}_{\mathbf{X}}C_{\mathbf{X}Y}}, \quad Z \sim \mathrm{N}(0, a\lambda \log e/2) \text{ independent of } \mathbf{X}$$

- $\lambda = 0$ (no rate constraint) $\Rightarrow$ linear regression (Farnia–Tse 2016)
- Straightforward estimate-compress scheme optimal:
  - Estimate: Compute MMSE estimate of Y given $\mathbf{X}$
  - Compress: Scale MMSE estimate and add Z to obtain $\hat{Y}$

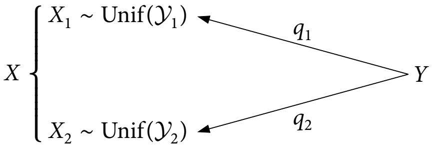


$|\mathcal{Y}_i| = k_i,\ i = 1, 2,\ q_2 = 1 - q_1$

- Let $\mathcal{Y} = \hat{\mathcal{Y}} = \mathcal{Y}_1 \cup \mathcal{Y}_2$, $\mathcal{Y}_1 \cap \mathcal{Y}_2 = \emptyset$, $|\mathcal{Y}_1| = k_1$, $|\mathcal{Y}_2| = k_2$; $l(y, \hat{y}) = \mathbb{1}_{\{\hat{y} \neq y\}}$;

  $\Gamma = \{P\}$, P: $X = (X_1, X_2) \sim \mathrm{Unif}[\mathcal{Y}_1 \times \mathcal{Y}_2]$, $Y = X_1$ w.p. q_1 or X_2 w.p. $q_2 = 1 - q_1$

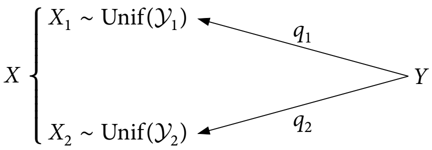
Classification example



$|\mathcal{Y}_i| = k_i,\ i = 1, 2,\ q_2 = 1 - q_1$

- If $q_1 > q_2$, MAP estimate is $\hat{Y} = X_1$

Classification example



$|\mathcal{Y}_i| = k_i,\ i = 1, 2,\ q_2 = 1 - q_1$

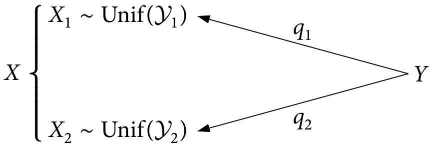
- Minimum risk-information cost: Let $a_1 = 2^{\lambda^{-1} q_1} + k_1 - 1$, $a_2 = 2^{\lambda^{-1} q_2} + k_2 - 1$,

$$\bar{L}^*_\lambda = 1 - \lambda \log \max \{a_1/k_1, a_2/k_2\},\ (*)$$

If $a_1/k_1 > a_2/k_2$, $\hat{Y} = \begin{cases} X_1 & \text{w.p. } a_1^{-1} 2^{\lambda^{-1} q_1}, \\ \sim \mathrm{Unif}(\mathcal{Y}_1 \setminus \{x_1\}) & \text{w.p. } a_1^{-1} \end{cases}$

If $a_1/k_1 \le a_2/k_2$, exchange 1 and 2 in above

Classification example



$|\mathcal{Y}_i| = k_i,\ i = 1, 2,\ q_2 = 1 - q_1$

- Minimum risk-information cost: Let $a_1 = 2^{\lambda^{-1} q_1} + k_1 - 1,\ a_2 = 2^{\lambda^{-1} q_2} + k_2 - 1,$

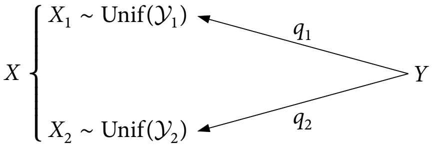
$$\bar{L}_\lambda^* = 1 - \lambda \log \max \{a_1/k_1, a_2/k_2\}, \quad (*)$$

$$\text{If } a_1/k_1 > a_2/k_2,\ \hat{Y} = \begin{cases} X_1 & \text{w.p. } a_1^{-1} 2^{\lambda^{-1} q_1}, \\ \sim \mathrm{Unif}(\mathcal{Y}_1 \setminus \{x_1\}) & \text{w.p. } a_1^{-1} \end{cases}$$

If $a_1/k_1 \le a_2/k_2$, exchange 1 and 2 in above

- Comparison to estimate-compress: If $q_1 > q_2$, MAP estimate $\hat{Y} = X_1$
  - Estimate-compress: Set $\hat{Y}$ to compressed X_1 or pick random $y \in \mathcal{Y}_2$,

$$\bar{L}_\lambda = 1 - \lambda \log \max \{a_1/k_1, 2^{\lambda^{-1} q_2 k_2^{-1}}\} \quad (**)$$



$|\mathcal{Y}_i| = k_i,\ i = 1, 2,\ q_2 = 1 - q_1$

- Minimum risk-information cost: Let $a_1 = 2^{\lambda^{-1} q_1} + k_1 - 1$, $a_2 = 2^{\lambda^{-1} q_2} + k_2 - 1$,

  $$\bar{L}^*_\lambda = 1 - \lambda \log \max\{a_1/k_1, a_2/k_2\}, \ (*)$$

  If $a_1/k_1 > a_2/k_2$, $\hat{Y} = \begin{cases} X_1 & \text{w.p. } a_1^{-1} 2^{\lambda^{-1} q_1}, \\ \sim \mathrm{Unif}(\mathcal{Y}_1 \setminus \{x_1\}) & \text{w.p. } a_1^{-1} \end{cases}$

  If $a_1/k_1 \le a_2/k_2$, exchange 1 and 2 in above

- Comparison to estimate-compress: If $q_1 > q_2$, MAP estimate $\hat{Y} = X_1$
  - Estimate-compress: Set $\hat{Y}$ to compressed X_1 or pick random $y \in \mathcal{Y}_2$,

    $$\bar{L}_\lambda = 1 - \lambda \log \max\{a_1/k_1, 2^{\lambda^{-1} q_2 k_2^{-1}}\} \ (**)$$

  - If $k_1 \gg k_2$, optimal scheme is to pick random $y \in \mathcal{Y}_2$ and $(**)$ can be >> $(*)$

Summary

- Strong functional representation lemma (SFRL)
  - $H(Y|Z)$ is between I and $I(X;Y) + \log I(X;Y)$
  - Poisson construction of Z, g

Summary

- Strong functional representation lemma (SFRL)
  - $H(Y|Z)$ is between I and $I(X;Y) + \log I(X;Y)$
  - Poisson construction of Z, g
- Applications of SFRL:
  - Channel simulation with common randomness
  - One-shot lossy compression
  - Minimax learning for distributed inference

    Estimate–compress is not optimal in general

Summary

- Strong functional representation lemma (SFRL)
  - $H(Y|Z)$ is between I and $I(X;Y) + \log I(X;Y)$
  - Poisson construction of Z, g
- Applications of SFRL:
  - Channel simulation with common randomness
  - One-shot lossy compression
  - Minimax learning for distributed inference

    Estimate–compress is not optimal in general
  - Other applications:

    Multiple description coding, Gray–Wyner system, Gelfand–Pinsker

Thank you!

References

Bennett, C. H., Devetak, I., Harrow, A. W., Shor, P. W., and Winter, A. (2014). The quantum reverse shannon theorem and resource tradeoffs for simulating quantum channels. *IEEE Trans. Info. Theory*, 60(5), 2926–2959.

Bennett, C. H., Shor, P. W., Smolin, J., and Thapliyal, A. V. (2002). Entanglement-assisted capacity of a quantum channel and the reverse Shannon theorem. *IEEE Trans. Info. Theory*, 48(10), 2637–2655.

Braverman, M. and Garg, A. (2014). Public vs private coin in bounded-round information. In *International Colloquium on Automata, Languages, and Programming*, Springer, pp. 502–513.

Cuff, P. (2013). Distributed channel synthesis. *IEEE Trans. Info. Theory*, 59(11), 7071–7096.

El Gamal, A. and Kim, Y.-H. (2011). *Network Information Theory*. Cambridge, Cambridge.

Elias, P. (1975). Universal codeword sets and representations of the integers. *IEEE Trans. Inf. Theory*, 21(2), 194–203.

Farnia, F. and Tse, D. (2016). A minimax approach to supervised learning. In *30th Conference on Neural Information Processing Systems*, pp. 4240–4248.

References (cont.)

Hajek, B. E. and Pursley, M. B. (1979). Evaluation of an achievable rate region for the broadcast channel. *IEEE Trans. Inf. Theory*, 25(1), 36–46.

Harsha, P., Jain, R., McAllester, D., and Radhakrishnan, J. (2010). The communication complexity of correlation. *IEEE Trans. Info. Theory*, 56(1), 438–449.

Kocaoglu, M., Dimakis, A., Vishwanath, S., and Hassibi, B. (2017). Entropic causality and greedy minimum entropy coupling. In *Proc. IEEE Symp. Info. Theory*, pp. 1465–1469.

Kostina, V., Polyanskiy, Y., and Verdú, S. (2015). Variable-length compression allowing errors. *IEEE Trans. Inf. Theory*, 61(8), 4316–4330.

Lee, J. and Raginsky, M. (2017). Minimax statistical learning and domain adaptation with Wasserstein distances. *arXiv preprint arXiv:1705.07815*.

Li, C. T. and El Gamal, A. (2018). Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11), 6967 – 6978.

Li, C. T., Wu, X., Özgür, A., and El Gamal, A. (2018). Minmax learning for remote prediction. In *IEEE International Symposium on Information Theory*, pp. 541–545.

References (cont.)

Namkoong, H. and Duchi, J. C. (2017). Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pp. 2975–2984.

Pinkston, J. (1967). *Encoding Independent Sample Information Sources*. Research Laboratory of Electronics, Massachusetts Inst. of Technology.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3), 379–423, 27(4), 623–656.

Song, E., Cuff, P., and Poor, H. (2016). The likelihood encoder for lossy compression. *IEEE Trans. Inf. Theory*, 62(4), 1836–1849.

Willems, F. M. J. and van der Meulen, E. C. (1985). The discrete memoryless multiple-access channel with cribbing encoders. *IEEE Trans. Inf. Theory*, 31(3), 313–327.

Zhang, Z., Yang, E., and Wei, V. (1997). The redundancy of source coding with a fidelity criterion. 1. known statistics. *IEEE Trans. Inf. Theory*, 43(1), 71–91.