The Rate Distortion Theory

Yonsei CS Theory Student Group Seminar

Information Theory Series, Week 5

Presented by Sungmin Kim on 24' Apr. 04

Motivation

High loss

Low loss

Topics to consider

- Definition of loss (=distortion)
- Relation between rate and loss
- Methods for computing optimal rate given loss

Rate and Distortion

Situation

X^n → encoder → $f_n(x^n) \in \{1, \ldots, 2^{nR}\}$ → decoder → $\hat{X}^n$

$X_1, \ldots, X_n \sim p(x)$ i.i.d.

- [Rate] Define rate R as the average number of bits per symbol for representing X^n.
  - Intuitively, we lose information as R decreases

Encode 4 pixels using a single pixel

$$R = 1/4$$

Rate and Distortion

Situation

$X^n \longrightarrow$ encoder $\xrightarrow{f_n(x^n) \in \{1, \dots, 2^{nR}\}}$ decoder $\longrightarrow \hat{X}^n$

$X_1, \dots, X_n \sim p(x)$ i.i.d.

- [Rate] Recall that, in the channel coding theorem, the definition of the rate R is

$$R = \frac{\log M}{n},$$

where M is the number of possible values of the channel input. Rearranging gives

$$M = 2^{nR}.$$

Rate and Distortion

Situation (lossy encoding)

X^n → encoder $\xrightarrow{f_n(x^n) \in \{1, \ldots, 2^{nR}\}}$ decoder → $\hat{X}^n$

$X_1, \ldots, X_n \sim p(x)$ i.i.d.

R is a parameter

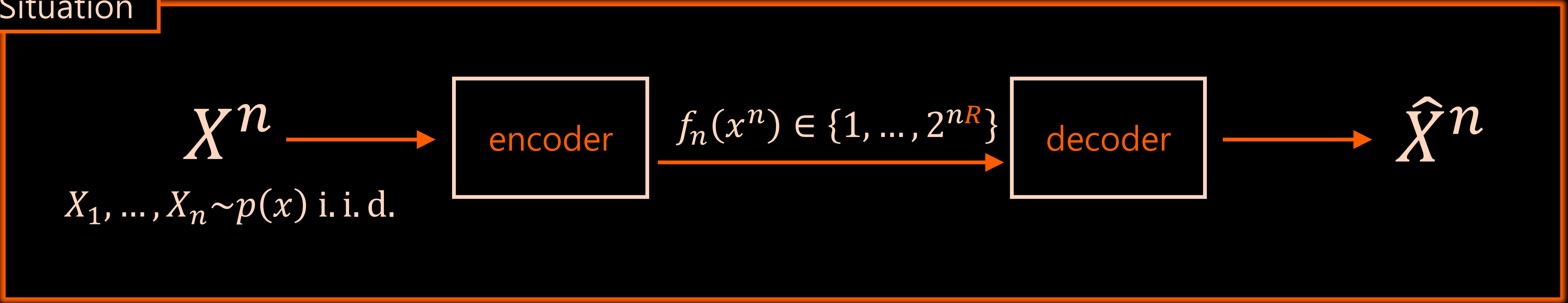
Situation (Channel Coding Thm)

encoder $\xrightarrow{X^n : \{1, \ldots, M\} \to \mathcal{X}}$ Discrete Memoryless Channel $\xrightarrow{Y^n : \mathcal{Y} \to \{1, \ldots, M\}}$ decoder

Fixed!

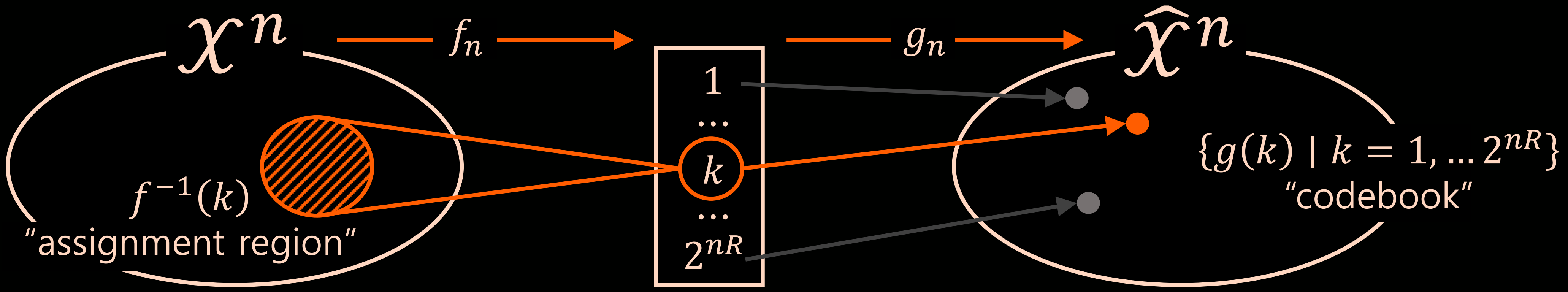
Rate and Distortion

Situation

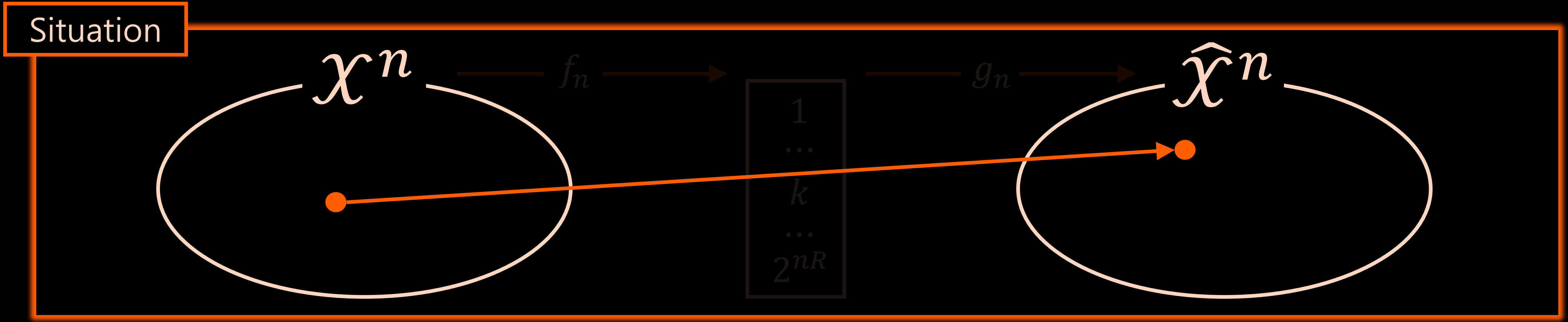


- [Distortion Code] Given a rate R, a function pair (f_n, g_n) where
  - the encoding function $f_n: \mathcal{X}^n \to \{1,2,\dots,2^{nR}\}$
  - the decoding function $g_n: \{1,2,\dots,2^{nR}\} \to \hat{\mathcal{X}}^n$

  is a $(2^{nR}, n)$-rate distortion code, i.e., the lossy encoding scheme.



Rate and Distortion



- [Distortion] Define distortion function $d: \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}^+$:
  - represents how different $\hat{X}$ is from X
  - options include Hamming distortion

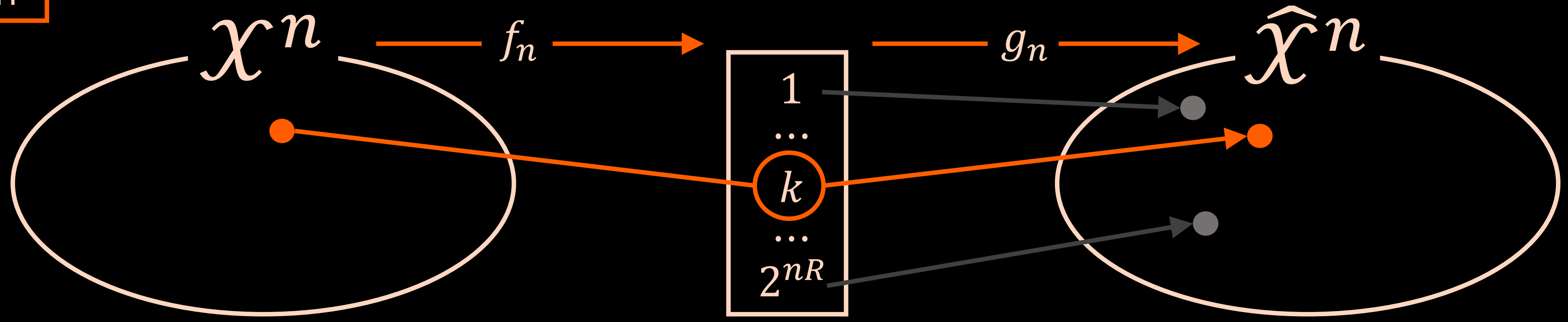
$$[d(x, \hat{x}) = 0] \Leftrightarrow [x = \hat{x}]$$

  - Or the squared-error distortion

$$d(x, \hat{x}) = (x - \hat{x})^2.$$

Rate and Distortion

Situation



- [Distortion between sequences] The distortion between x^n and $\hat{x}^n$ is

$$d(x^n, \hat{x}^n) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i).$$

- [Distortion of a code] The distortion D of a $(2^{nR}, n)$-rate distortion code (f_n, g_n) is

$$D = \mathbb{E}\left[d\left(X^n, g_n\left(f_n(X^n)\right)\right)\right],$$

the expected distortion over all X^n values.

Rate and Distortion

Situation

$X^n \longrightarrow$ encoder $\xrightarrow{f_n(x^n) \in \{1, \ldots, 2^{nR}\}}$ decoder $\longrightarrow \hat{X}^n$

$X_1, \ldots, X_n \sim p(x)$ i.i.d.

- [Achievability] The rate-distortion pair (R, D) is achievable if there exists a sequence of $(2^{nR}, n)$-rate distortion codes (f_n, g_n) such that

$$\lim_{n\to\infty} \mathbb{E}\left[d\left(X^n, g_n(f_n(X^n))\right)\right] \le D.$$

- [Rate Distortion Function] The rate distortion function $R(D)$ gives the infimum of rates R such that (R, D) is in the closure of the set of achievable rate distortion pairs.

The Information Rate Distortion Function

Def The information rate distortion function $R^{(I)}(D)$ is defined as the following:

$$\min_{p(x,\hat{x}):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x})\leq D} I(X;\hat{X})$$

i.e., the minimum mutual information over all joint distributions $p(x,\hat{x})$ with total distortion at most D.

Thm 10.2.1 The minimum achievable rate at distortion D is exactly

$$R(D) = R^{(I)}(D).$$

- [part 1] $R \geq R^{(I)}(D)$ for any $(2^{nR}, n)$-rate distortion code with distortion $\leq D$.
- [part 2] $\left(R^{(I)}(D), D\right)$ is achievable.

The Minimum Achievable Rate for Distortion

Prove $R \geq R^{(I)}(D)$ for any $(2^{nR}, n)$-rate distortion code with distortion $\leq D$.

Lem $R^{(I)}(D)$ is convex and non-increasing in D.

- [non-increasing] if D increases, more joint distributions $p(x, \hat{x})$ should be considered;

$$\min_{p(x,\hat{x}):\sum_{(x,\hat{x})} p(x)p(\hat{x}|x)d(x,\hat{x})\leq D} I(X;\hat{X})$$

Thus, $R^{(I)}(D)$ is non-increasing in D.

The Minimum Achievable Rate for Distortion

Prove $R \geq R^{(I)}(D)$ for any $(2^{nR}, n)$-rate distortion code with distortion $\leq D$.

Lem $R^{(I)}(D)$ is convex and non-increasing in D.

[convexity] First, rewrite

$$D = \mathbb{E}\left[d\left(X^n, g_n\left(f_n(X^n)\right)\right)\right] = \sum_{(x,\hat{x})} p(x^n, \hat{x}^n) d(x^n, \hat{x}^n)$$

i.e., D is linear in $p(\hat{x}^n \mid x^n)$.

Now, consider (R_1, D_1) and (R_2, D_2), both on the rate distortion curve.

Let $p_1(x, \hat{x}) = p(x)p_1(\hat{x} \mid x)$ and $p_2(x, \hat{x}) = p(x)p_2(\hat{x} \mid x)$ be their distributions, resp.

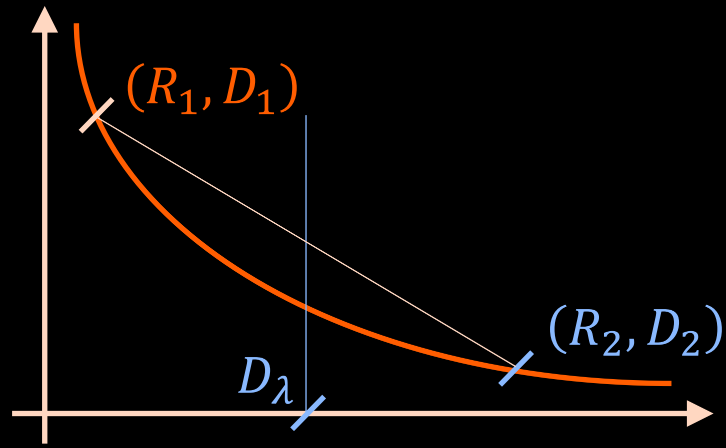
Let $p_\lambda = \lambda p_1 + (1-\lambda)p_2$ and we have $D_\lambda = \lambda D_1 + (1-\lambda)D_2$ by linearity in $p(\hat{x}^n \mid x^n)$.

The Minimum Achievable Rate for Distortion

Prove $R \geq R^{(I)}(D)$ for any $(2^{nR}, n)$-rate distortion code with distortion $\leq D$.

Lem $R^{(I)}(D)$ is convex and non-increasing in D.

- [convexity-cont.] Recall that $I(X;\hat{X})$ is convex (Thm. 2.7.4)



$$\begin{aligned} R^{(I)}(D_\lambda) &\leq I_{p_\lambda}(X;\hat{X}) \\ &\leq \lambda I_{p_1}(X;\hat{X}) + (1-\lambda) I_{p_2}(X;\hat{X}) \\ &= \lambda R(D_1) + (1-\lambda) R(D_2) \end{aligned}$$

Thus, $R^{(I)}(D)$ is convex in D.

The Minimum Achievable Rate for Distortion

Prove: $R \geq R^{(I)}(D)$ for any $(2^{nR}, n)$-rate distortion code with distortion $\leq D$.

- Follow the series of inequalities:

$$nR \geq H\left(f_n(X^n)\right) \quad \text{[Property of } H\text{] } H(X) \leq \log|\mathcal{X}|$$

$$\geq H\left(f_n(X^n)\right) - H(f_n(X^n) \mid X^n)$$

$$= I\left(X^n; f_n(X^n)\right) \quad \text{[Definition of mutual information]}$$

$$\geq I\left(X^n; \hat{X}^n\right) \quad \text{[Data processing inequality]}$$

$$= H(X^n) - H\left(X^n \mid \hat{X}^n\right)$$

$$= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H\left(X_i \mid \hat{X}^n, X_{i-1}, \dots, X_1\right) \quad \text{[}X_i\text{'s independent] [Chain rule]}$$

The Minimum Achievable Rate for Distortion

Prove $R \geq R^{(I)}(D)$ for any $(2^{nR}, n)$-rate distortion code with distortion $\leq D$.

- Follow the series of inequalities:

$$
\begin{aligned}
nR &\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H\left(X_i \mid \hat{X}^n, X_{i-1}, \ldots, X_1\right) \\
&\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H\left(X_i \mid \hat{X}_i\right) \quad \text{······ [Conditioning reduces entropy]} \\
&= \sum_{i=1}^{n} I\left(X_i; \hat{X}_i\right)
\end{aligned}
$$

The Minimum Achievable Rate for Distortion

$\boxed{\text{Prove}}$ $R \geq R^{(I)}(D)$ for any $(2^{nR}, n)$-rate distortion code with distortion $\leq D$.

- Follow the series of inequalities:

$$nR \geq \sum_{i=1}^{n} I(X_i; \hat{X}_i)$$

$$\geq \sum_{i=1}^{n} R^{(I)}\left(\mathbb{E}\left[d(X_i, \hat{X}_i)\right]\right) \quad \cdots\cdots \text{[Definition of } R^{(I)}(D)\text{]}$$

$$\geq nR^{(I)}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[d(X_i, \hat{X}_i)\right]\right)\right) \cdots \text{[Convexity of } R^{(I)}(D)\text{]}$$

The Minimum Achievable Rate for Distortion

Prove $R \geq R^{(I)}(D)$ for any $(2^{nR}, n)$-rate distortion code with distortion $\leq D$.

- Follow the series of inequalities:

$$nR \geq nR^{(I)}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left[d\left(X_i, \hat{X}_i\right)\right]\right)\right)$$

$$= nR^{(I)}\left(\mathbb{E}\left[d\left(X, \hat{X}\right)\right]\right) \quad \text{[Definition of } d(X, \hat{X})\text{]}$$

$$\geq nR^{(I)}(D) \quad \text{[}R^{(I)}(D) \text{ non-increasing]} \quad \text{[}\mathbb{E}[d(X, \hat{X})] \leq D \text{ from condition]}$$

- Therefore, we have $R \geq R^{(I)}(D)$ for any $(2^{nR}, n)$-rate distortion code with distortion $\leq D$.

The Minimum Achievable Rate for Distortion

Prove $\left(R^{(I)}(D), D\right)$ is achievable.

Claim For any $\delta > 0$, there exists a rate distortion code with rate R and distortion $\leq D + \delta$.

Proof Technical; uses the distortion ϵ-typicality to bound probabilities

- $\left|-\frac{1}{n}\log p(x^n) - H(X)\right| < \epsilon,$
- $\left|-\frac{1}{n}\log p(\hat{x}^n) - H(\hat{X})\right| < \epsilon,$
- $\left|-\frac{1}{n}\log p(x^n, \hat{x}^n) - H(X, \hat{X})\right| < \epsilon,$
- $\left|d(x^n, \hat{x}^n) - \mathbb{E}\left[d(X, \hat{X})\right]\right| < \epsilon.$

Characterizing the Rate Distortion Function

Prob Compute the rate distortion function

$$R(D) = \min_{q(\hat{x}|x):\sum_{(x,\hat{x})} p(x)q(\hat{x}|x)d(x,\hat{x})\le D} I(X;\hat{X}).$$

- Recall that $I(X;\hat{X})$ is convex; the problem is a minimization of a convex function over the convex set of all $q(x \mid \hat{x}) \ge 0$ satisfying the constraints
  - $\sum_{\hat{x}} q(\hat{x} \mid x) = 1$ for all x,
  - $\sum_{(x,\hat{x})} p(x)q(\hat{x} \mid x)d(x,\hat{x}) \le D$.
- Reformulate the problem using Lagrange multipliers and we get…

Characterizing the Rate Distortion Function

Prob | Optimize the following functional:

$$J(q) = \sum_x \sum_{\hat{x}} p(x) q(\hat{x} \mid x) \log \frac{q(\hat{x} \mid x)}{\sum_x p(x) q(\hat{x} \mid x)}$$

$$+\lambda \sum_x \sum_{\hat{x}} p(x) q(\hat{x} \mid x) d(x, \hat{x})$$

$$+\sum_x \nu(x) \sum_{\hat{x}} q(\hat{x} \mid x)$$ .... [conditional probability]

- We want to know $q(\hat{x}) = \sum_x p(x) q(\hat{x} \mid x)$ values for all $\hat{x} \in \hat{\mathcal{X}}$.

Characterizing the Rate Distortion Function

Sol An optimal solution gives us for all $\hat{x} \in \hat{\mathcal{X}}$,

$$\sum_x \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}'} q(\hat{x}')e^{-\lambda d(x,\hat{x}')}} = 1.$$

- From the definition of distortion, we know

$$\sum_{(x,\hat{x})} p(x)q(\hat{x} \mid x)d(x,\hat{x}).$$

- Now we can solve for $q(\hat{x})$ and λ.
- However, we usually have constraints on $q(\hat{x})$...

Computing the Rate Distortion Function

Lem Let $p(x)p(y \mid x)$ be a given joint distribution. Then,

$$D\big(p(x)p(y \mid x)||p(x)r^*(y)\big) = \min_{r(y)} D\big(p(x)p(y \mid x)||p(x)r(y)\big)$$

where $r^*(y) = \sum_x p(x)p(y \mid x)$.

Proof Subtract LHS from $D\big(p(x)p(y \mid x)||p(x)r(y)\big)$ for any $r(y)$ to get ≥ 0.

Computing the Rate Distortion Function

Rewrite the rate distortion function (again)

$$R(D) = \min_{q(\hat{x}|x):\sum_{(x,\hat{x})} p(x)q(\hat{x}|x)d(x,\hat{x})\le D} I(X;\hat{X})$$

$$= \min_{q(\hat{x}|x):\sum_{(x,\hat{x})} p(x)q(\hat{x}|x)d(x,\hat{x})\le D} D\big(p(x)q(\hat{x}\mid x)\,||\,p(x)q(\hat{x})\big)$$

Recall that $q(\hat{x}) = \sum_x q(\hat{x},x) = \sum_x p(x)q(\hat{x}\mid x) = r^*(y)$ from the prev. lemma.

$$= \underbrace{\min_{r(\hat{x})}}_{\text{Minimize B}} \underbrace{\min_{q(\hat{x}|x):\sum_{(x,\hat{x})} p(x)q(\hat{x}|x)d(x,\hat{x})\le D}}_{\text{Minimize A}} D\big(p(x)q(\hat{x}\mid x)\,||\,p(x)r(\hat{x})\big)$$

Computing the Rate Distortion Function

Alg [Blahut-Arimoto]

Given: distortion D, input distribution $p(x)$ where X_i's are i.i.d. sampled from

Goal: Compute the conditional probability $q(\hat{x} \mid x)$ that minimizes $R(D)$.

- Choose initial λ and $|\hat{\mathcal{X}}|$ values $r(\hat{x})$.
- Repeat until convergence:
  - Minimize A by solving for all $(x, \hat{x}) \in \mathcal{X} \times \hat{\mathcal{X}}$

$$q(\hat{x} \mid x) = \frac{r(\hat{x}) e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{x}} r(\hat{x}) e^{-\lambda d(x,\hat{x})}}$$

Optimization w/ Lagrangian multiplier

  - Minimize B by computing for all $\hat{x} \in \hat{\mathcal{X}}$

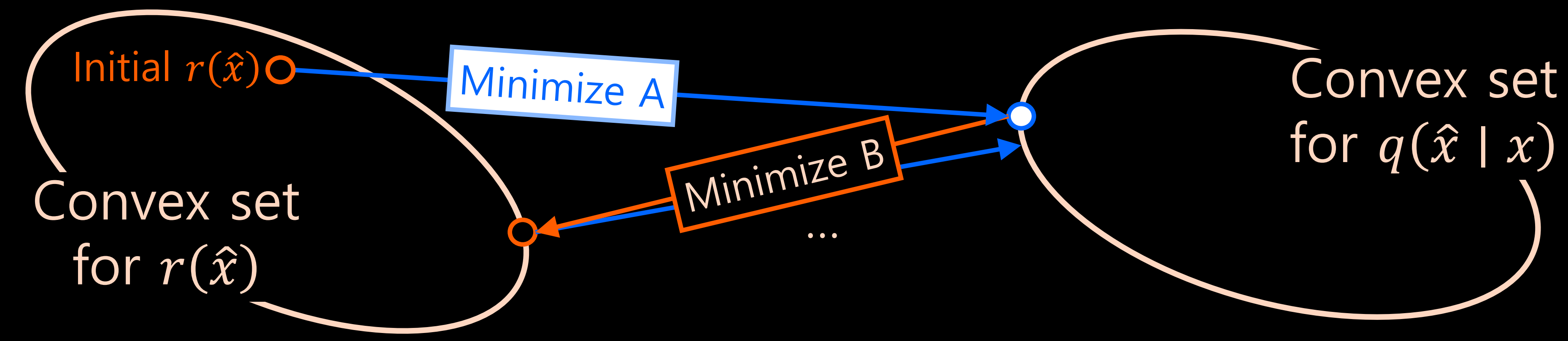
$$r(\hat{x}) = \sum_{x} p(x) q(\hat{x} \mid x)$$

Lemma

Computing the Rate Distortion Function

Thm [Csiszár] The Blahut-Arimoto algorithm converges to a distribution that gives rate $R(D)$.

Situation



Remark

- Higher λ means less compression
- Similar algorithm used for computing channel capacity

Some Final Remarks

Thm 10.4.1 For a discrete memoryless channel with capacity C, distortion rate D is achievable if and only if $C > R(D)$.

Thm 10.3.1 The rate distortion function of a Bernoulli(p) source w/ Hamming distortion is given by

$$R(D) = \begin{cases} H(p) - H(D), & 0 \leq D \leq \min\{p, 1-p\}, \\ 0, & D > \min\{p, 1-p\}. \end{cases}$$

Thm 10.3.2 The rate distortion function of a $\mathcal{N}(0, \sigma^2)$ source w/ squared-error distortion is given by

$$R(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2, \\ 0, & D > \sigma^2. \end{cases}$$

Things to Discuss

- ☐ No consensus on a "good" distortion metric for human perception
- ☐ What if the input is not i.i.d. sampled from $\mathcal{X}$?

References

- Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*, 2nd edition. Wiley, 2006.