Lecture 13

Water-filling Intuition + Transform Coding

Announcements

Quiz Q1

You have been given following joint probability distribution table for $$(X,Y)$$ on binary alphabets:

P(X=x,Y=y)	y = 0	y = 1
x = 0	0.5	0
x = 1	0.25	0.25

1.1 Calculate the joint entropy $H(X,Y)$.

$H(X,Y) = \sum_{x,y} P(X=x,Y=y) \log_2 \frac{1}{P(X=x,Y=y)} = 1.5$.

1.2 Calculate the mutual information $I(X;Y)$.

$I(X;Y) = H(X) + H(Y) - H(X,Y) = H_b(0.5) + H_b(0.75) - 1.5 = 0.31$

Quiz Q2

Consider a uniformly distributed source on alphabet $\{0, 1, 2\}$.

You have been asked to lossily compress this source under MSE (mean square error) distortion and have been asked to calculate the rate distortion function $R(D)$ for a given distortion value D.

2.1 What is $R(D = 0)$?

$R(D = 0) = H(X) = \log_2 3$

2.2 What is $R(D = 1)$?

$R(D = 1) = 0$!, since we can always send 1 and achieve distortion $D(X_i, \hat{X}_i) <= 1$.

Quiz Q3

For a $Ber(1/2)$ source with Hamming distortion, we saw in class that $R(D) = 1 - H_b(D)$, where $H_b(p)$ is entropy of a binary random variable with probability p. Which of the following are correct?
(Choose all that apply)

[ ] There exists a scheme working on large block sizes achieving distortion D and rate < $1 - H_b(D)$.
[ ] There exists a scheme working on large block sizes achieving distortion D and rate > $1 - H_b(D)$.
[ ] There exists a scheme working on large block sizes achieving distortion D and rate arbitrarily close to $1 - H_b(D)$.
[ ] There exists a scheme working on single symbols at a time (block size = 1) achieving distortion D and rate arbitrarily close to $1 - H_b(D)$.

Recap

1. Learnt about Mutual Information

   Let X, Y be two random variables with joint distribution $p(x, y)$. Then we define the mutual information between X, Y as:

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

Recap

2. Learnt about (Shannon's) Rate-Distortion theory.

   Let $X_1, X_2, \ldots$ be data generated i.i.d. Then, the optimal rate $R(D)$ for a given maximum distortion D is:

   $$R(D) = \min_{\mathbb{E}d(X,Y) \leq D} I(X;Y)$$

   where the expectation in the minimum is over distributions $q(x,y) = p(x)q(y|x)$, where $q(y|x)$ are any arbitrary conditional distributions.

Recap

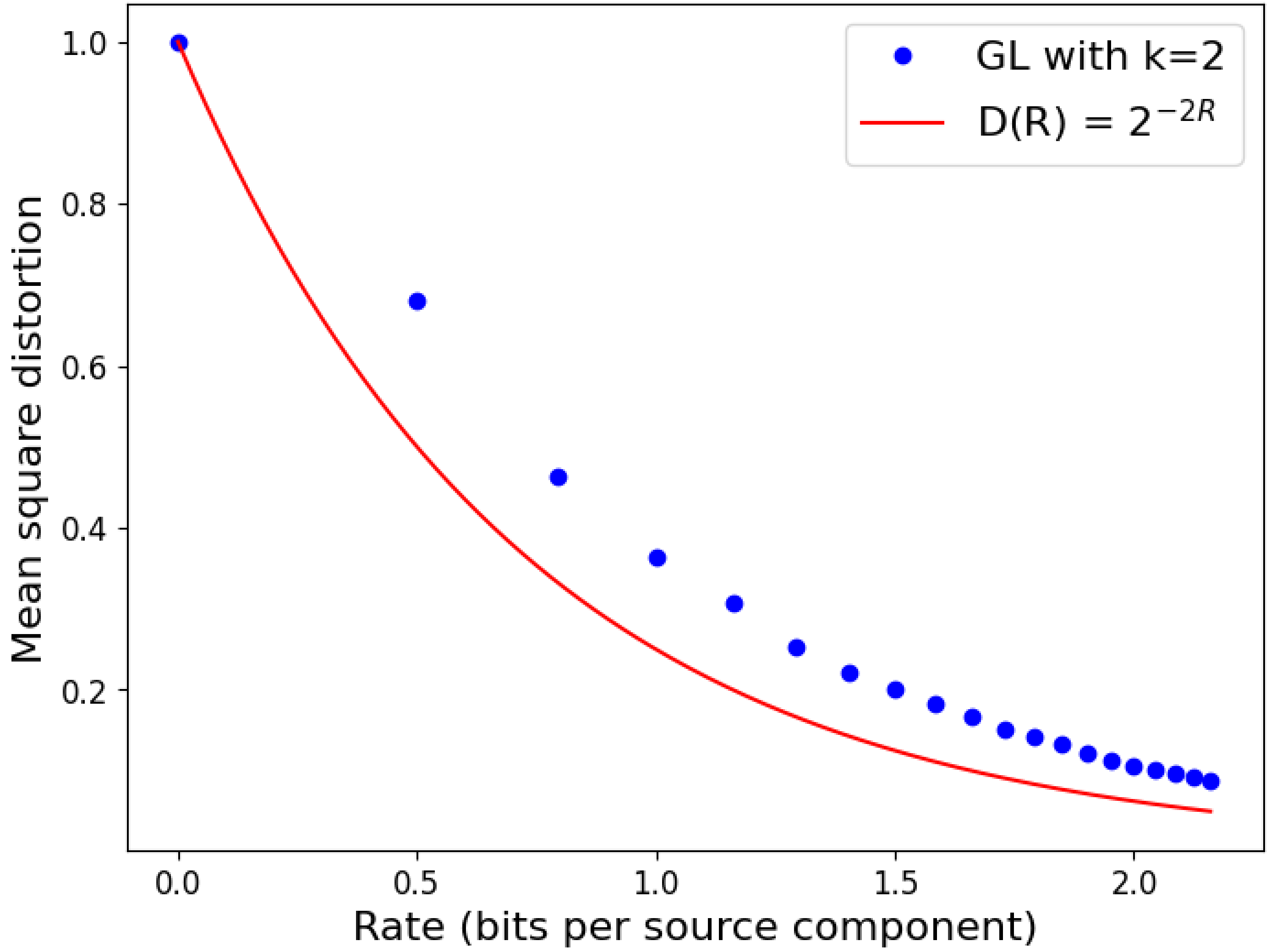
3. Saw example for Gaussian Sources under MSE distortion.

   Let $X \sim \mathcal{N}(0, \sigma^2)$, i.e. the data samples $X_1, X_2, \ldots$ are distributed as unit gaussians. Also, lets consider the distortion to be the mean square distortion: $d(x,y) = (x-y)^2$ i.e the mse distortion. Then:

$$R(D) = \begin{cases} \frac{1}{2}\log_2 \frac{\sigma^2}{D} & 0 \le D \le \sigma^2 \\ 0 & D > \sigma^2 \end{cases}$$

   Also denoted by $R_G(\sigma^2, D) = \left(\frac{1}{2}\log_2 \frac{\sigma^2}{D}\right)_+$

Recap: Performance



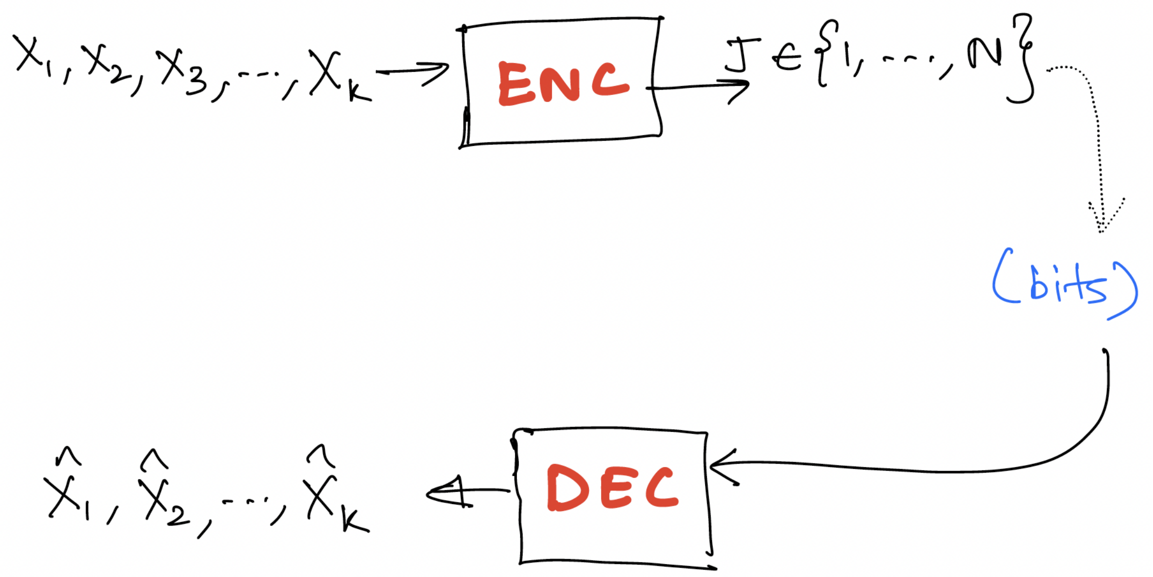
Thumb-rule for Lossy Compression

> **Thumb-rule**: For a given distortion measure, allocate more bits to the components with higher variance.

Today

1. Water-filling intuition for correlated gaussian sources
2. Learn about Transform Coding

Lossy Compression Problem Formulation



The two metrics for lossy compression are:

- Rate R $= \frac{log N}{k}$ bits/source component
- Distortion D $= d(X^k, \hat{X}^k) = \frac{1}{k}\sum_{i=1}^{k} d(X_i, \hat{X}_i)$

Generalization of Shannon's RD Theorem

Let $X_1, X_2, \ldots$ be data generated $\underline{I.I.D.}$. Then, the optimal rate $R(D)$ for a given maximum distortion D is:

$$R(D) = \min_{\mathbb{E}d(X,Y) \leq D} I(X;Y)$$

This is also referred to as *memoryless* sources.

But what if the data is $\underline{correlated}$?

Generalization of Shannon's RD Theorem

Consider source X^n and reconstruction $\hat{X}^n$. Then,

$$R(X^n, D) = min_{E[d(X^n,\hat{X}^n)]\leq D} \frac{1}{n} I(X^n; \hat{X}^n)$$

i.e. Shannon's RD theorem generalizes to correlated sources as well.

- Just like $R(X, D)$ was the analog of entropy of X, $R(X^n, D)$ is the analog of entropy of the n-tuple.

Generalization of Shannon's RD Theorem

Consider source X^n and reconstruction $\hat{X}^n$. Let $\mathbf{X} = X_1, X_2, X_3, \ldots$ define a stationary stochastic process. Then,

$$R(\mathbf{X}, D) = \lim_{n \to \infty} R(X^n, D)$$

- $R(\mathbf{X}, D)$ is the analog of entropy rate of the n-tuple.
  - can show this limit exists for stationary sources.

> the best you can do for stationary processes, in the limit of encoding arbitrarily many symbols in a block, is $R(\mathbf{X}, D)$

Example: Gaussian Source, $k = 2$

- Let $X_1 \sim N(0, \sigma_1^2)$, $X_2 \sim N(0, \sigma_2^2)$ be independent random variables.
- Then, $X^2 = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ is a 2-dimensional random vector.
- Notation: $R(X^2, D) = R_G\left(\begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \end{bmatrix}, D\right)$.

It can be shown that:

$$R_G\left(\begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \end{bmatrix}, D\right) = min_{\frac{1}{2}(D_1+D_2)\leq D} \frac{1}{2}[R_G(\sigma_1^2, D_1) + R_G(\sigma_1^2, D_2)]$$

i.e. we can greedily optimize independently over each component of the vector, ensuring that the total distortion is less than D.

Example: Gaussian Source, $k = 2$

$$R_G\left(\begin{bmatrix}\sigma_1^2\\ \sigma_2^2\end{bmatrix}, D\right) = min_{\frac{1}{2}(D_1+D_2)\leq D}\frac{1}{2}[R_G(\sigma_1^2, D_1) + R_G(\sigma_1^2, D_2)]$$

$$= min_{\frac{1}{2}(D_1+D_2)\leq D}\ \frac{1}{2}\left[\left(\frac{1}{2}\log\frac{\sigma_1^2}{D_1}\right)_+ + \left(\frac{1}{2}\log\frac{\sigma_2^2}{D_2}\right)_+\right]$$

Can be solved using convex optimization techniques (solving KKT conditions). We will look into the answer for some intuition.

Example: Gaussian Source; Intuition

WLOG: assume $\sigma_1^2 \leq \sigma_2^2$

$$R_G\left(\begin{bmatrix}\sigma_1^2 \\ \sigma_2^2\end{bmatrix}, D\right) = min_{\frac{1}{2}(D_1+D_2)\leq D}\ \frac{1}{2}\left[\left(\frac{1}{2}\log\frac{\sigma_1^2}{D_1}\right)_+ + \left(\frac{1}{2}\log\frac{\sigma_2^2}{D_2}\right)_+\right]$$

Quiz-1: Should I ever allow $D_1 > \sigma_1^2$?

Example: Gaussian Source; Intuition

WLOG: assume $\sigma_1^2 \leq \sigma_2^2$

$$R_G\left(\begin{bmatrix}\sigma_1^2\\ \sigma_2^2\end{bmatrix}, D\right) = min_{\frac{1}{2}(D_1+D_2)\leq D}\ \frac{1}{2}\left[\left(\frac{1}{2}\log\frac{\sigma_1^2}{D_1}\right)_+ + \left(\frac{1}{2}\log\frac{\sigma_2^2}{D_2}\right)_+\right]$$

Quiz-1: Should I ever allow $D_1 > \sigma_1^2$?

Quiz-2: What is $R(D_1)$ if $D_1 > \sigma_1^2$?

Example: Gaussian Source; Intuition

WLOG: assume $\sigma_1^2 \leq \sigma_2^2$

$$R_G\left(\begin{bmatrix}\sigma_1^2 \\ \sigma_2^2\end{bmatrix}, D\right) = min_{\frac{1}{2}(D_1+D_2)\leq D}\ \frac{1}{2}\left[\left(\frac{1}{2}\log\frac{\sigma_1^2}{D_1}\right)_+ + \left(\frac{1}{2}\log\frac{\sigma_2^2}{D_2}\right)_+\right]$$

Quiz-3: What is $R(D)$ if $D > \frac{\sigma_1^2+\sigma_2^2}{2}$

Example: Gaussian Source; Solution

Let R(D) curve be parameterized by θ, i.e. $R(\theta), D(\theta)$. Then, solution to the optimization problem

$$R_G\left(\begin{bmatrix}\sigma_1^2\\ \sigma_2^2\end{bmatrix}, D\right) = min_{\frac{1}{2}(D_1+D_2)\leq D}\ \frac{1}{2}\left[\left(\frac{1}{2}\log\frac{\sigma_1^2}{D_1}\right)_+ + \left(\frac{1}{2}\log\frac{\sigma_2^2}{D_2}\right)_+\right]$$

is given by:

- $D_i = \min\{\theta, \sigma_i^2\}$ for $i = 1, 2$; and $\frac{1}{2}(D_1 + D_2) = D$.
- $R = \frac{1}{2}\left[\left(\frac{1}{2}\log\frac{\sigma_1^2}{D_1}\right)_+ + \left(\frac{1}{2}\log\frac{\sigma_2^2}{D_2}\right)_+\right]$

i.e. we can find θ which satisfies the first condition, giving us the R(D) curve as $R(\theta), D(\theta)$.

Example: Gaussian Source; Water-filling Intuition

3 cases (WLOG: assume $\sigma_1^2 \leq \sigma_2^2$):

1. $D < \sigma_1^2$ and $D < \sigma_2^2$
2. $\sigma_1^2 < D < \sigma_2^2$
3. $D > \frac{\sigma_1^2+\sigma_2^2}{2}$

Example: Gaussian Source; Water-filling Intuition

One of the main ideas in lossy-compression, recall thumb-rule!

> **Thumb-rule**: For a given distortion measure, allocate more bits to the components with higher variance.

For a block of 2 components, we can allocate more bits to the component with higher variance.
This is the **water-filling intuition**.

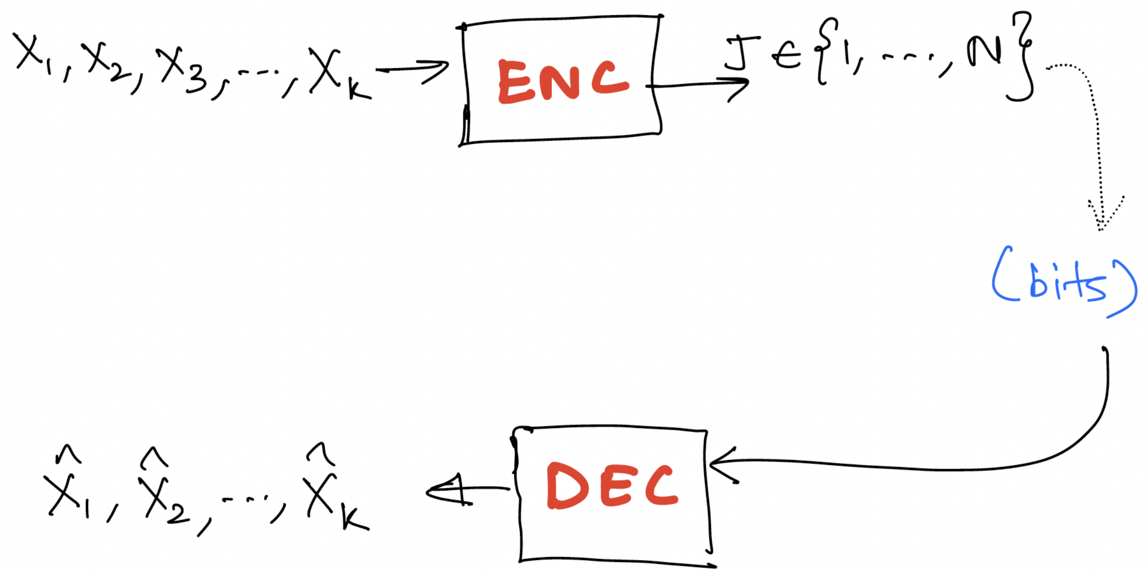
Onto Transform Coding: A Few Comments

- We looked into an example of uncorrelated gaussian sources, and saw that we can use water-filling intuition to selectively allocate bits to different components.
- This generalizes beautifully to *correlated gaussian processes* as well (see notes).
- But in general, we will have correlated non-gaussian sources, and we will need to do something more sophisticated.

> **Transform Coding**: Transform the source to a different domain to allow for decorrelated components with different variances. Then, use water-filling intuition to selectively allocate bits to different components of the transformed source.

Transform Coding

(recall) Lossy compression problem formulation:



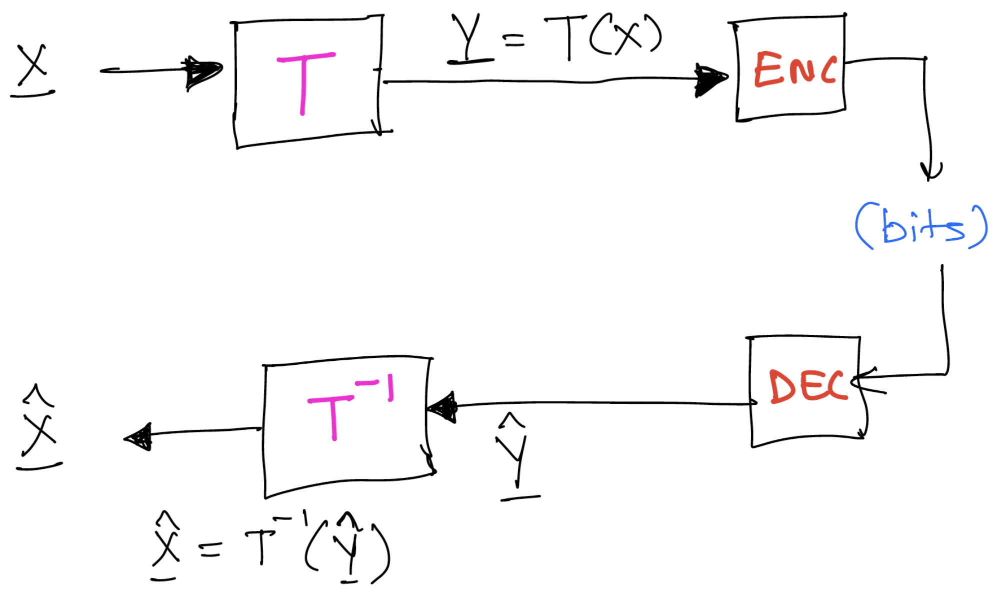
The two metrics for lossy compression are:

- Rate R $= \frac{logN}{k}$ bits/source component
- Distortion D $= d(X^k, \hat{X}^k) = \frac{1}{k}\sum_{i=1}^{k} d(X_i, \hat{X}_i)$

Transform Coding

Notation: $X^k = (X_1, \ldots, X_k)$ as $\underline{X}$. Therefore, $\underline{X} \in \mathbb{R}^k$ (vector).

- Convert $\underline{X}$ to $\underline{Y} = T(\underline{X})$, for this class assume T is linear (matrix)
- Need that T should be invertible
- We can use scalar or vector quantization on $\underline{Y}$ to get $\hat{\underline{Y}}$



Transform Coding

Why transform coding?

- **Decorrelation**: X can be correlated, aim to de-correlate it
  - allows for efficient coding of $\underline{Y}$ e.g. using scalar quantization instead of vector quantization
- **Energy compaction**: more *energy* in first few components of $\underline{Y}$ than in the last few
  - allows for allocating bits to different components of $\underline{Y}$ in a more-efficient manner (recall: water-filling!)

This gives us criterion as to how we would like to choose T.

We will look into a specific transform T which is an *orthonormal matrix*.

Linear Algebra Review: Orthogonal Matrices

Consider $Y = AX$ (matrix-vector product). If A is orthonormal (denoted by U), then:

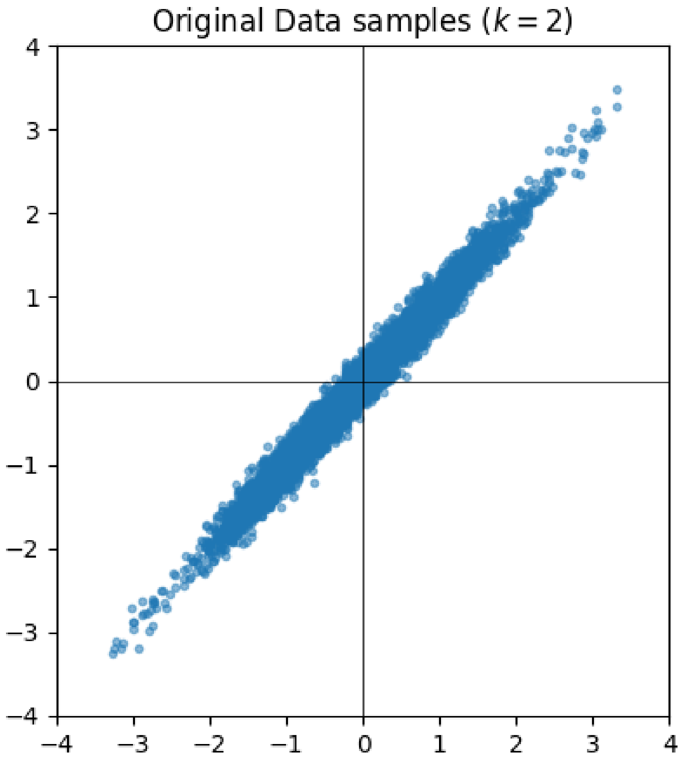
- $U^T U = I$ (orthonormality)
- Square of the Euclidean norm, also called *energy* in the signal, is preserved under transform:
  - $||Y||^2 = Y^T Y = X^T U^T U X = X^T X = ||X||^2$
  - This is also called the Parseval's theorem in context of Fourier transform.
  - This says that the energy in transform domain matches the energy in the original.
- The transform preserves Euclidian distances between points, i.e.,
  - if $Y_1 = UX_1$ and $Y_2 = UX_2$, then $||Y_1 - Y_2||^2 = ||X_1 - X_2||^2$.
  - Allows us to do analysis for MSE distortion!
  - $D_{MSE} = \mathbb{E}||X - \hat{X}||^2 = \mathbb{E}||Y - \hat{Y}||^2$

Linear Algebra Review: Eigenvalue Decomposition/Decorrelation

- Any symmetric matrix A can be decomposed as $A = U\Lambda U^T$, where U is orthonormal and Λ is diagonal.
- U is the matrix of (normalized) eigenvectors of A and Λ is the matrix of eigenvalues of A.
- U is orthonormal, i.e., $U^TU = I$.
- We can use this to get de-correlated components of X by using $Y = U^TX$, i.e. $T = U^T$.
    - Let covariance matrix of X be $\Sigma = \mathbb{E}[XX^T]$.
    - We can apply eigenvalue decomposition to get $\Sigma = U\Lambda U^T$.
    - Then, $Y = U^TX$ is de-correlated, i.e., $\mathbb{E}[YY^T] = \mathbb{E}[U^TXX^TU] = U^T\mathbb{E}[XX^T]U = U^T\Sigma U = \Lambda$.

Decorrelation Example

Example: consider a source $X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\mathcal{N}(0,\sigma^2)$, $X_0 \sim \mathcal{N}(0,\sigma^2)$.
We will work with blocks of 2, i.e. $k = 2$.



Decorrelation Example

Example: consider a source $X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\mathcal{N}(0,\sigma^2)$, $X_0 \sim \mathcal{N}(0,\sigma^2)$. We will work with blocks of 2, i.e. $k = 2$.

Quiz-4: What is the 2×2 covariance matrix Σ of X?

HINT: your sequence is stationary!

$$\Sigma = \mathbb{E}\left[\begin{bmatrix} X_i - \mathbb{E}X_i \\ X_{i+1} - \mathbb{E}X_{i+1} \end{bmatrix} \begin{bmatrix} X_i - \mathbb{E}X_i & X_{i+1} - \mathbb{E}X_{i+1} \end{bmatrix}\right]$$

Decorrelation Example

Example: consider a source $X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\mathcal{N}(0, \sigma^2)$, $X_0 \sim \mathcal{N}(0, \sigma^2)$. We will work with blocks of 2, i.e. $k = 2$.

Quiz-4: What is the 2×2 covariance matrix Σ of X?

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \sigma^2$$

Decorrelation Example

Example: consider a source $X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\mathcal{N}(0,\sigma^2)$, $X_0 \sim \mathcal{N}(0,\sigma^2)$. We will work with blocks of 2, i.e. $k = 2$.

Can show that the eigenvalues of Σ are
- $\lambda_1 = (1+\rho)\sigma^2$ and $\lambda_2 = (1-\rho)\sigma^2$
- corresponding eigenvectors are $u_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$ and $u_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$.

Quiz-5: What is the eigenvalue-based transform at block-size $k = 2$ and transformed components Y?

Decorrelation Example

Example: consider a source $X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\mathcal{N}(0,\sigma^2)$, $X_0 \sim \mathcal{N}(0,\sigma^2)$. We will work with blocks of 2, i.e. $k=2$.

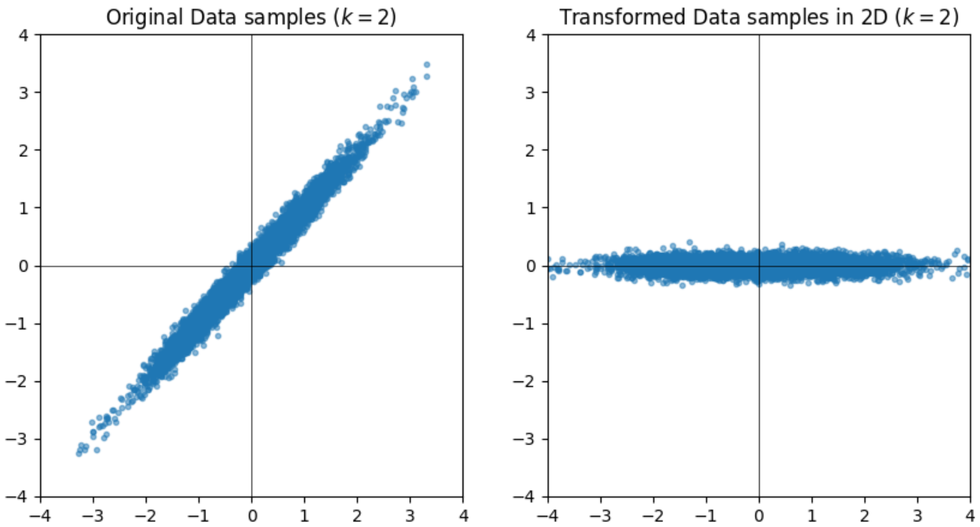
Quiz-5: What is the eigenvalue-based transform at block-size $k=2$, transformed components Y?

$T = U^T = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}$ and therefore $Y = TX = \frac{1}{\sqrt{2}}\begin{bmatrix}X_i + X_{i+1}\\ X_i - X_{i+1}\end{bmatrix}$

Decorrelation Example

Example: consider a source $X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\mathcal{N}(0,\sigma^2)$, $X_0 \sim \mathcal{N}(0,\sigma^2)$. We will work with blocks of 2, i.e. $k = 2$.

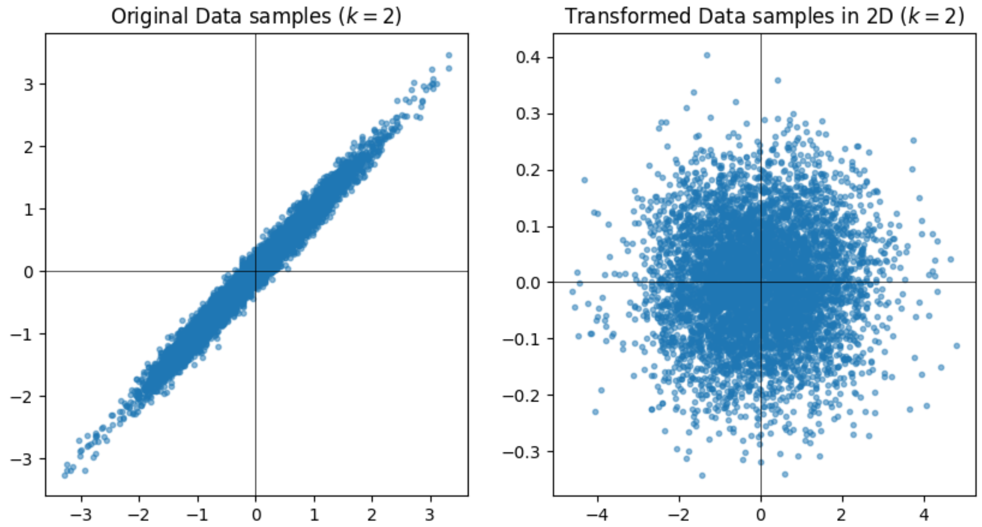
$$Y = TX = \frac{1}{\sqrt{2}}\begin{bmatrix} X_i + X_{i+1} \\ X_i - X_{i+1} \end{bmatrix}$$



Quiz-6: What is the 2×2 covariance matrix Σ of Y?

Decorrelation Example

Example: consider a source $X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\mathcal{N}(0,\sigma^2)$, $X_0 \sim \mathcal{N}(0,\sigma^2)$. We will work with blocks of 2, i.e. $k=2$.



Quiz-6: What is the 2×2 covariance matrix Σ_Y of Y?

$\Sigma_Y = \begin{bmatrix} (1+\rho) & 0 \\ 0 & (1-\rho) \end{bmatrix} \sigma^2$, i.e. Y_1 and Y_2 are uncorrelated!

Moreover, the variances of Y_1 and Y_2 are such that Y_1 has higher variance than Y_2. This is the energy compaction property of the transform. (recall: water-filling!)

Karhunen-Loeve Transform (KLT)

- We looked into what is called the **Karhunen-Loeve Transform (KLT)** in signal processing.
- The KLT is the eigenvalue-based linear transform.
- The KLT is the *optimal* transform for a given covariance matrix Σ (without proof).
    - By optimal, we mean it in the sense that it maximally reduces the correlation between the transformed components.
    - The components have the property that they are uncorrelated and ordered in decreasing order of variance.
- Useful for many applications: often used for data compression, dimensionality reduction, and feature extraction in various fields, including image and signal processing.

Transform Coding + KLT

- We looked into one specific transform, the KLT, which is an orthonormal matrix and allows us to decorrelate the data.

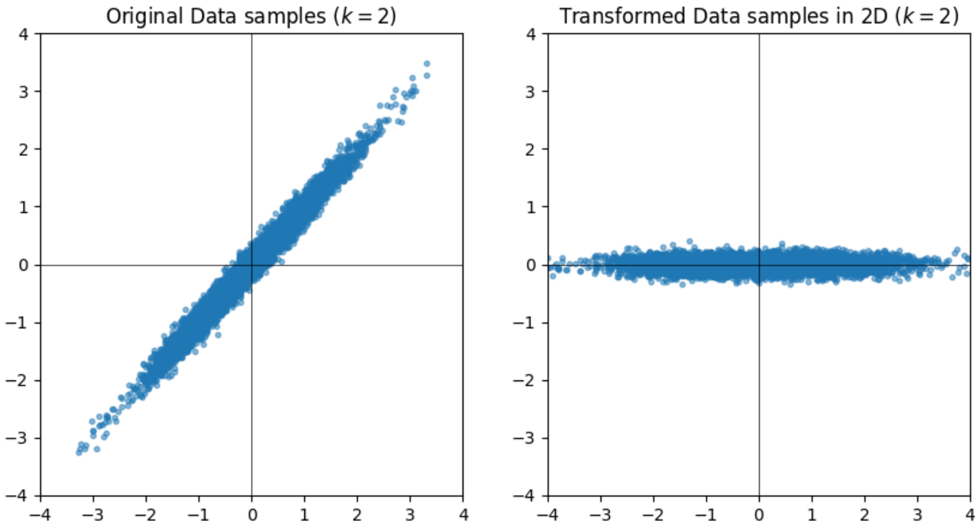
Quiz-7: How does this allow better lossy-compression of X?

Transform Coding + KLT

- We looked into one specific transform, the KLT, which is an orthonormal matrix and allows us to decorrelate the data.

Quiz-7: How does this allow better lossy-compression of X?

For MSE distortion, we can allocate bits to the transformed components Y in a more-efficient manner, i.e., allocate more bits to the components with higher energy. (recall: thumb-rule!)



Transform Coding Notebook

https://colab.research.google.com/drive/1ZcnjIco0HEbiTQWvcpiPYA9HbtfB829x?usp=sharing

Transform Coding Performance on our Example

Example: consider a source $X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\mathcal{N}(0,\sigma^2)$

```
==================================================
Processing rho: 0.9
==================================================
Vector Quantization Experiment
==================================================
[VQ][Bit per symbol: 1][Block Size: 2]Rate: 1.0, Distortion: 0.163
[VQ][Bit per symbol: 1][Block Size: 4]Rate: 1.0, Distortion: 0.095
==================================================
TC Vector Quantization Experiment
==================================================
[TC_VQ][Bit per symbol: 1][Block Size: 2][Bitrate Split: [1, 1]]Rate: 1.0, Distortion: 0.276
[TC_VQ][Bit per symbol: 1][Block Size: 2][Bitrate Split: [0, 2]]Rate: 1.0, Distortion: 0.970
[TC_VQ][Bit per symbol: 1][Block Size: 2][Bitrate Split: [2, 0]]Rate: 1.0, Distortion: 0.122
================================================
```

Transform Coding Performance on our Example

Example: consider a source $X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\mathcal{N}(0,\sigma^2)$

```
==================================================
Processing rho: 0.99
==================================================
Vector Quantization Experiment
==================================================
[VQ][Bit per symbol: 1][Block Size: 2]Rate: 1.0, Distortion: 0.107
[VQ][Bit per symbol: 1][Block Size: 4]Rate: 1.0, Distortion: 0.020
==================================================
TC Vector Quantization Experiment
==================================================
[TC_VQ][Bit per symbol: 1][Block Size: 2][Bitrate Split: [1, 1]]Rate: 1.0, Distortion: 0.204
[TC_VQ][Bit per symbol: 1][Block Size: 2][Bitrate Split: [0, 2]]Rate: 1.0, Distortion: 0.890
[TC_VQ][Bit per symbol: 1][Block Size: 2][Bitrate Split: [2, 0]]Rate: 1.0, Distortion: 0.030
================================================
```

Transform Coding Performance on our Example

Example: consider a source $X_n = \rho X_{n-1} + \sqrt{1-\rho^2}\mathcal{N}(0, \sigma^2)$

```
==================================================
Processing rho: 0.5
==================================================
Vector Quantization Experiment
==================================================
[VQ][Bit per symbol: 1][Block Size: 2]Rate: 1.0, Distortion: 0.305
[VQ][Bit per symbol: 1][Block Size: 4]Rate: 1.0, Distortion: 0.271
==================================================
TC Vector Quantization Experiment
==================================================
[TC_VQ][Bit per symbol: 1][Block Size: 2][Bitrate Split: [1, 1]]Rate: 1.0, Distortion: 0.374
[TC_VQ][Bit per symbol: 1][Block Size: 2][Bitrate Split: [0, 2]]Rate: 1.0, Distortion: 0.786
[TC_VQ][Bit per symbol: 1][Block Size: 2][Bitrate Split: [2, 0]]Rate: 1.0, Distortion: 0.343
================================================
```

Transform Coding + KLT: Issues

Quiz-8: Can you think of any issues with doing KLT in practice?

Transform Coding + KLT: Issues

Quiz-8: Can you think of any issues with doing KLT in practice?

Ans:

- KLT is dependent on statistics of input data X!
  - KLT is optimal for a given covariance matrix Σ.
  - In practice, we do not know Σ and need to estimate it from data.
  - Moreover, data in real-life is not stationary, i.e., statistics change over time. Need to re-estimate Σ.
  - Therefore, in practice, KLT is computationally expensive!

Next class we will see other *fixed* orthonormal transforms which are more practical such as DCT, DFT, wavelets, etc.