Neural Networks Optimally Compress the Sawbridge

Aaron Wagner
Cornell University

Johannes Ballé
Google Research

Classical Rate-Distortion Theory

Classical Rate-Distortion Theory

- The classical approach to lossy compression subject to an MSE constraint:

Classical Rate-Distortion Theory

- The classical approach to lossy compression subject to an MSE constraint:

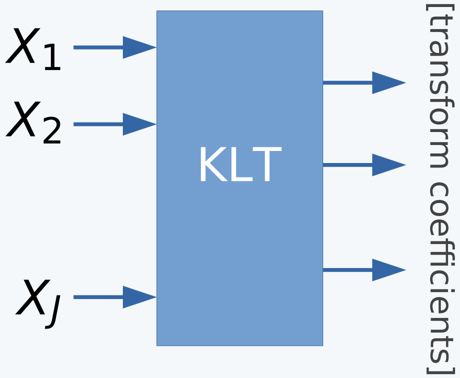
X_1

X_2

X_J

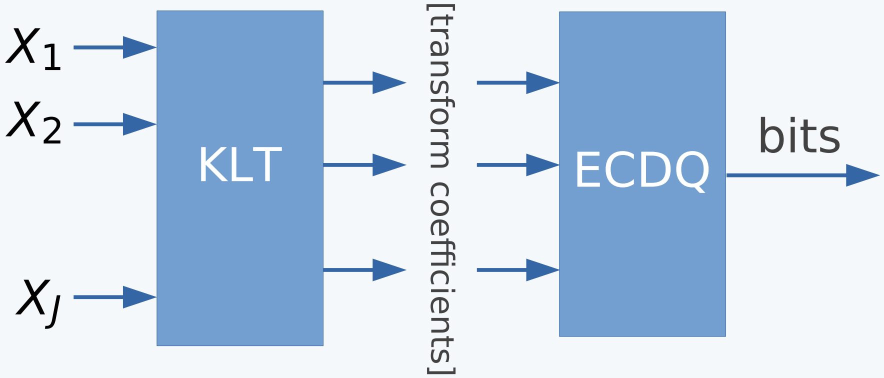
Classical Rate-Distortion Theory

- The classical approach to lossy compression subject to an MSE constraint:



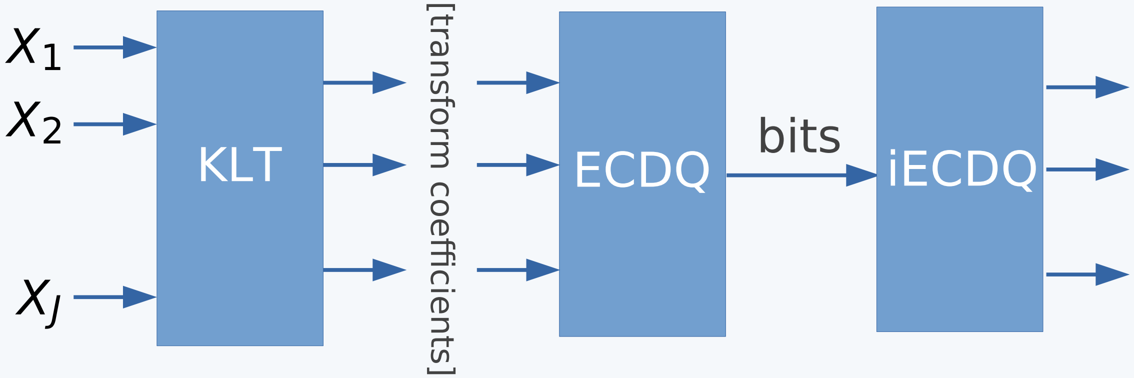
Classical Rate-Distortion Theory

- The classical approach to lossy compression subject to an MSE constraint:



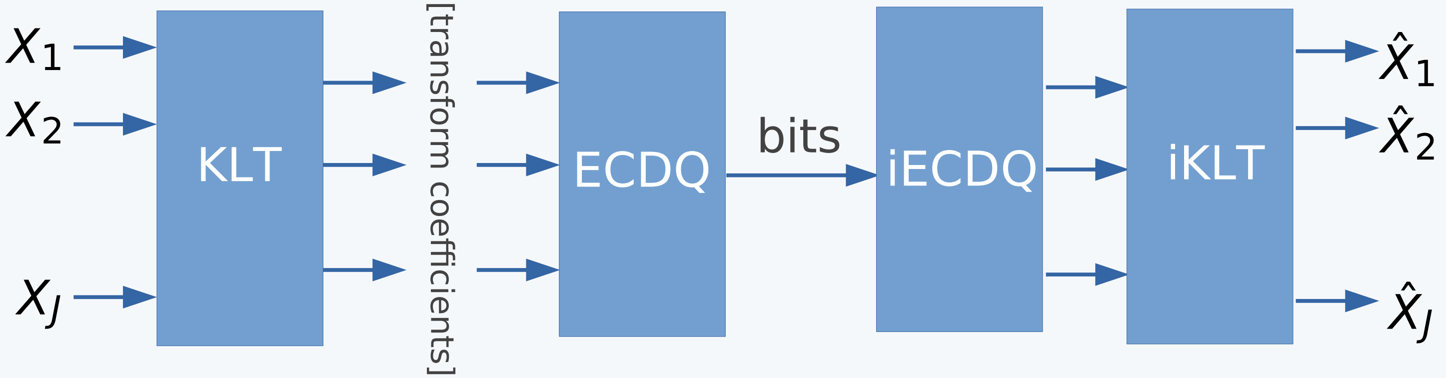
Classical Rate-Distortion Theory

- The classical approach to lossy compression subject to an MSE constraint:



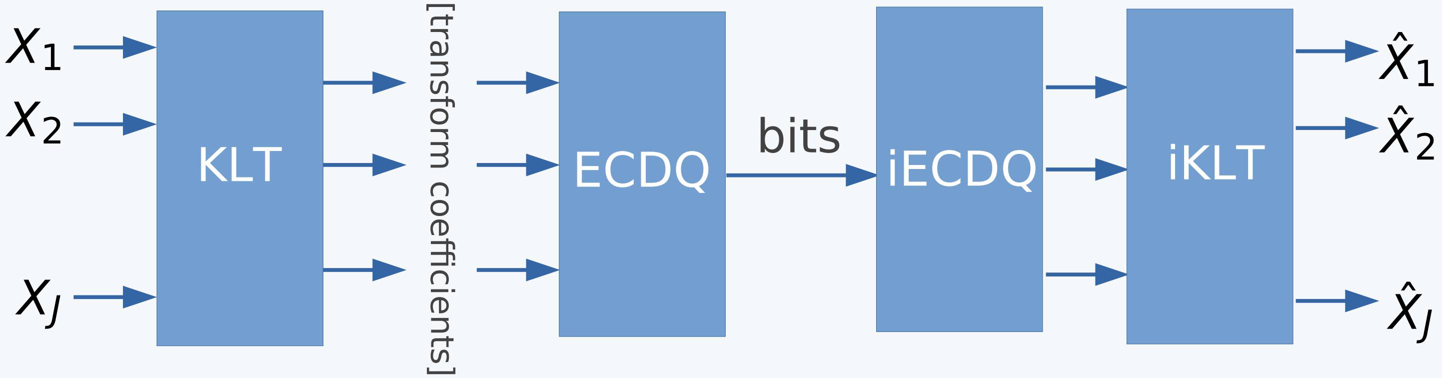
Classical Rate-Distortion Theory

- The classical approach to lossy compression subject to an MSE constraint:



Classical Rate-Distortion Theory

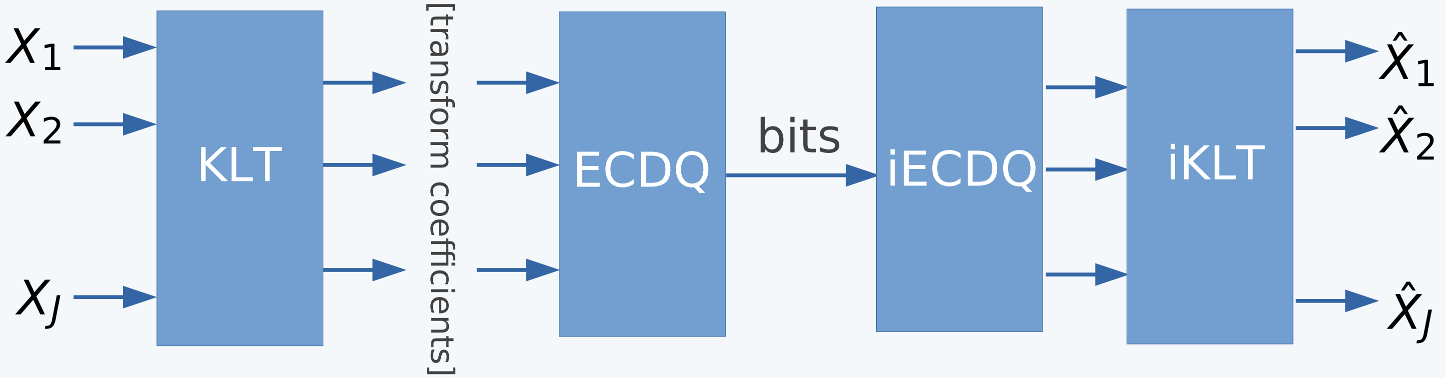
- The classical approach to lossy compression subject to an MSE constraint:



- ... is within .255 bits/sample of optimal for stationary Gaussian sources at high rates.

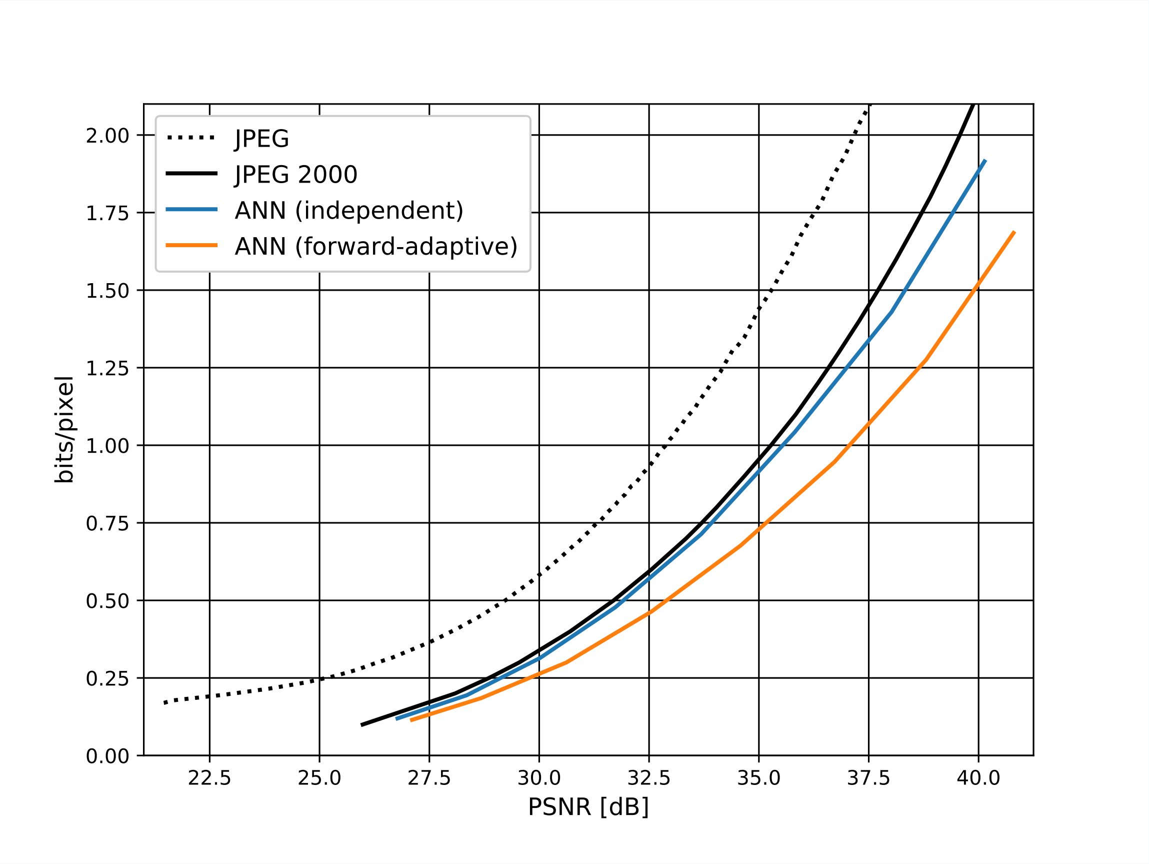
Classical Rate-Distortion Theory

- The classical approach to lossy compression subject to an MSE constraint:



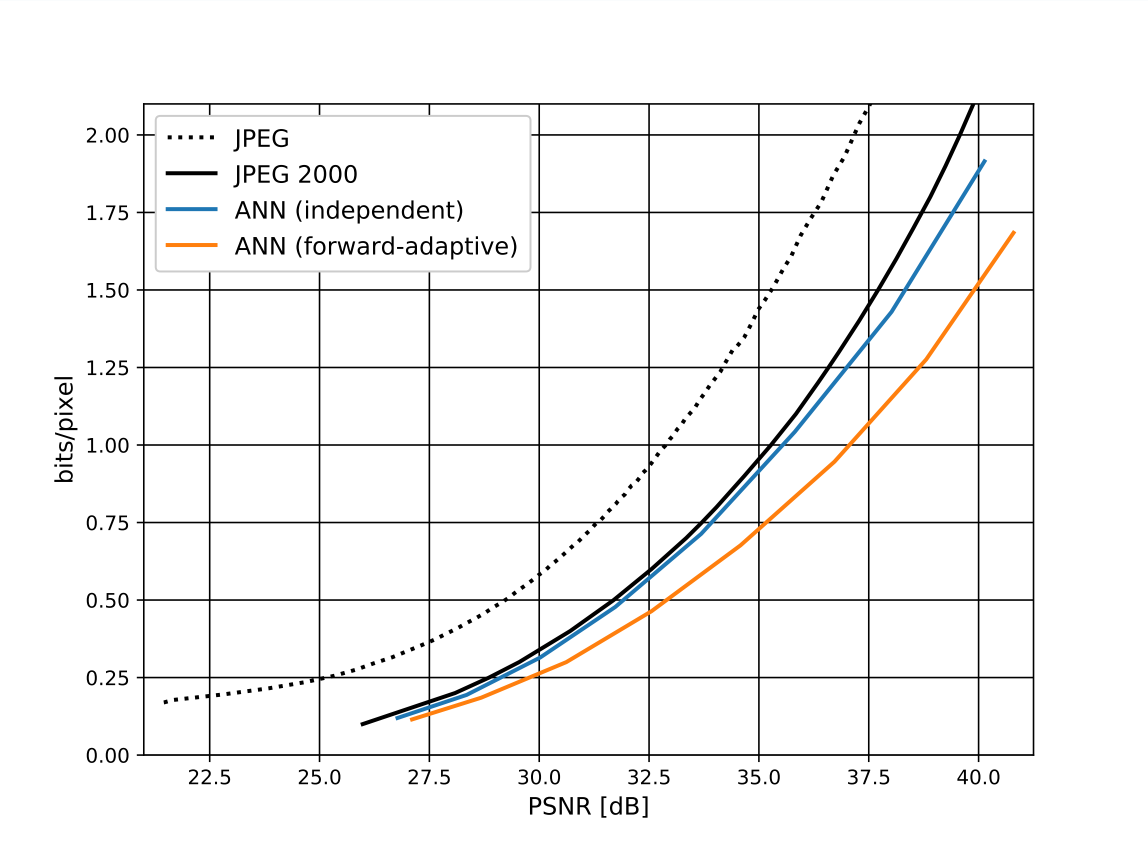
- ... is within .255 bits/sample of optimal for stationary Gaussian sources at high rates.
- Images are assumed to live on a low-D manifold with large linear span.

Success of Artificial Neural Network (ANN) Compressors



[from Ballé *et al.* ‘18]

Success of Artificial Neural Network (ANN) Compressors

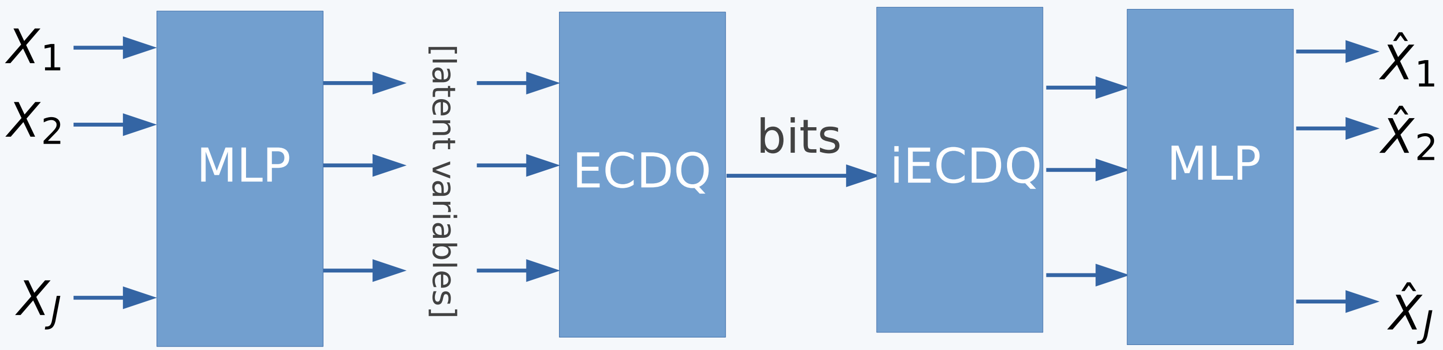


[from Ballé *et al.* ‘18]

[See also Ballé *et al.* ‘16, Theis *et al.* ‘17, Rippel and Bourdev, ‘17, Toderici *et al.* ‘17]

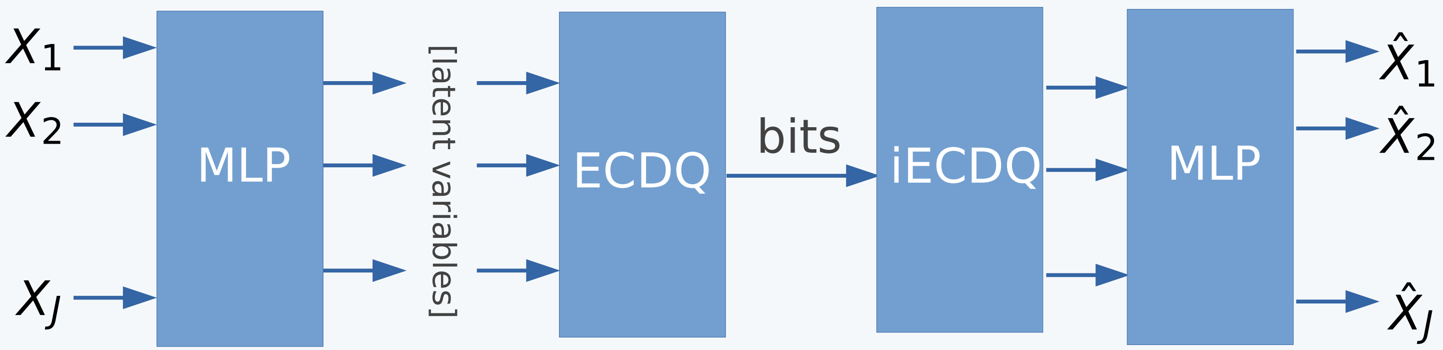
Nonlinear Transform Coding (NTC)

- A high-level description:



Nonlinear Transform Coding (NTC)

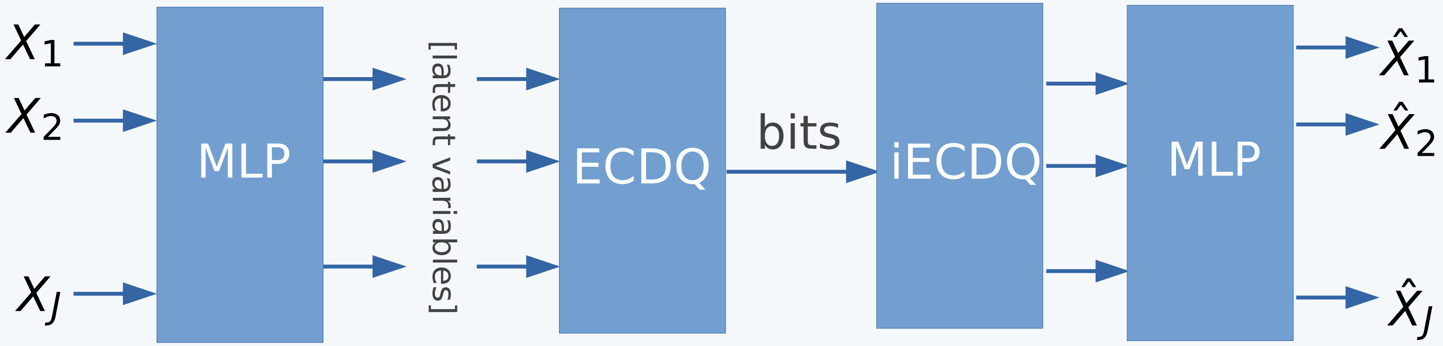
- A high-level description:



- Versus classical approach:

Nonlinear Transform Coding (NTC)

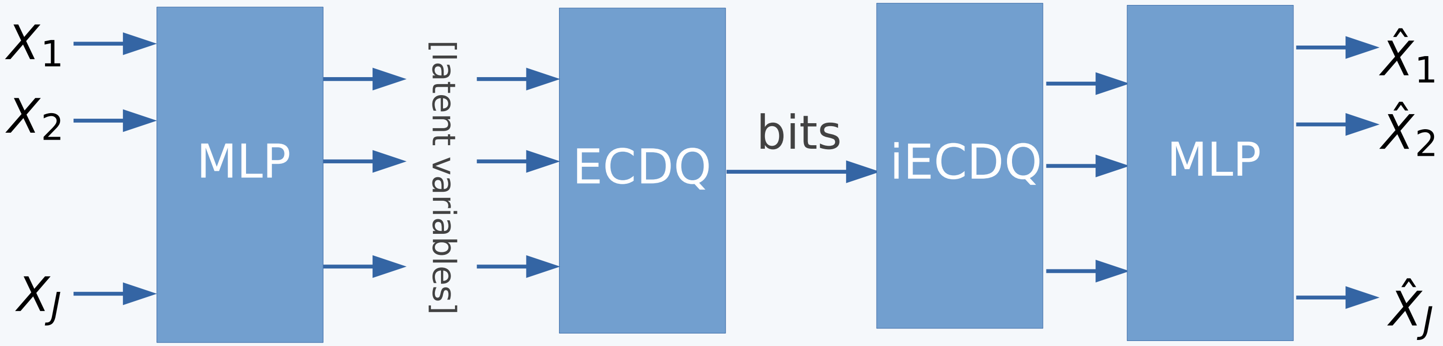
- A high-level description:



- Versus classical approach:
  - Learned (data-driven) vs. modeled

Nonlinear Transform Coding (NTC)

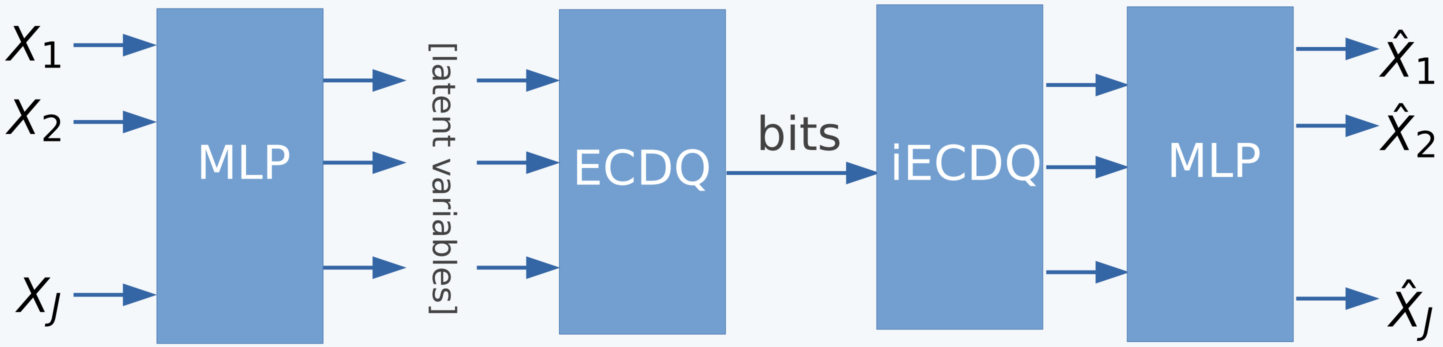
- A high-level description:



- Versus classical approach:
  - Learned (data-driven) vs. modeled
  - Multilevel perceptron (MLP) vs. Karhunen-Loève Transform (KLT)

Nonlinear Transform Coding (NTC)

- A high-level description:



- Versus classical approach:
  - Learned (data-driven) vs. modeled
  - Multilevel perceptron (MLP) vs. Karhunen-Loève Transform (KLT)
- Is this scheme optimal for some image-like source models?

The Sawbridge

The Sawbridge

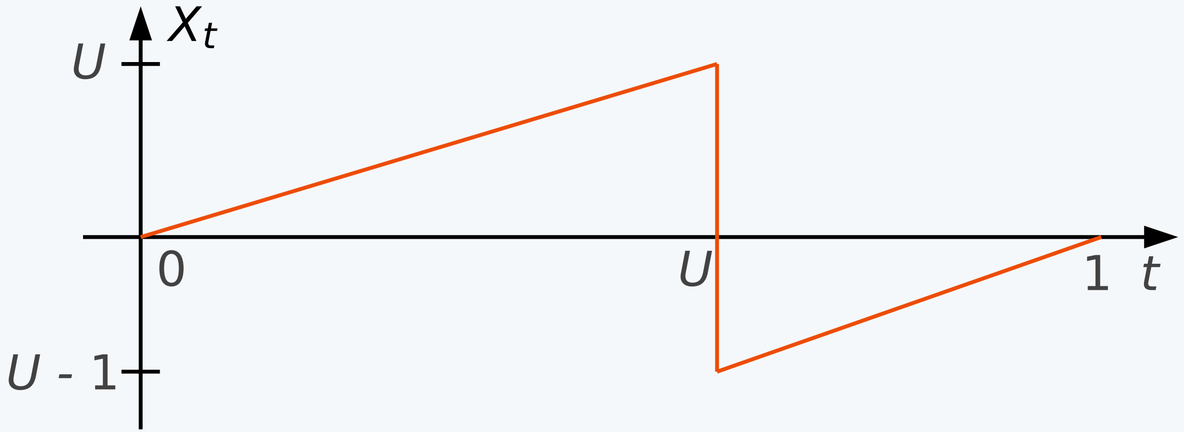
- Let U be unif[0,1] and

$$X(t) = t - \mathbf{1}(U \leq t) \quad t \in [0, 1]$$

The Sawbridge

- Let U be unif[0,1] and

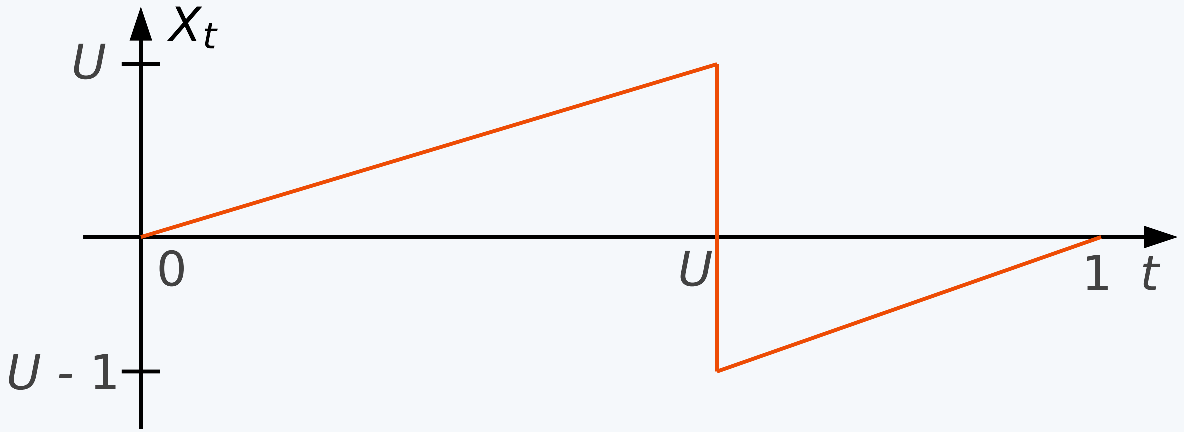
$$X(t) = t - \mathbf{1}(U \leq t) \quad t \in [0, 1]$$



The Sawbridge

- Let U be unif[0,1] and

$$X(t) = t - \mathbf{1}(U \le t) \quad t \in [0, 1]$$

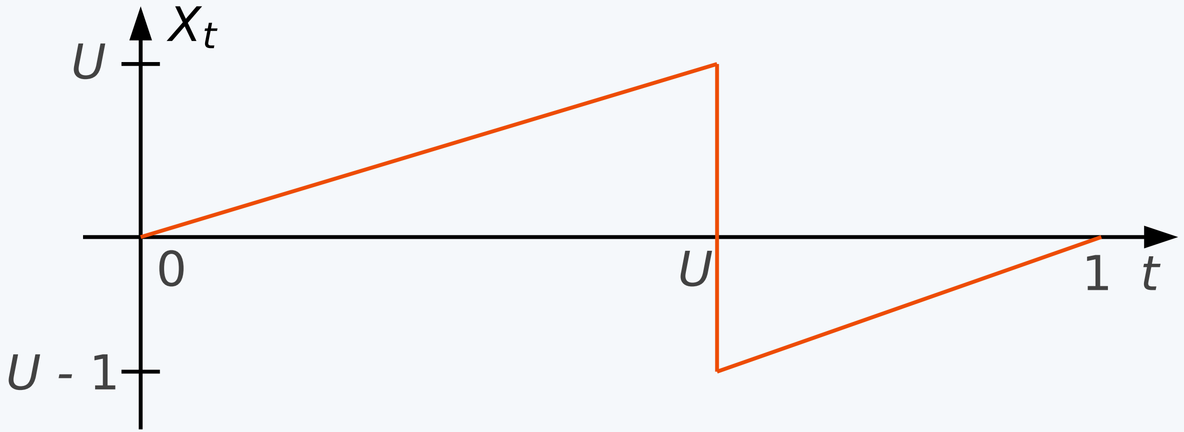


- Donoho, Vetterli, DeVore, and Daubechies ‘98 (“Ramp”); Meyer ‘92

The Sawbridge

- Let U be unif[0,1] and

$$X(t) = t - \mathbf{1}(U \leq t) \quad t \in [0, 1]$$

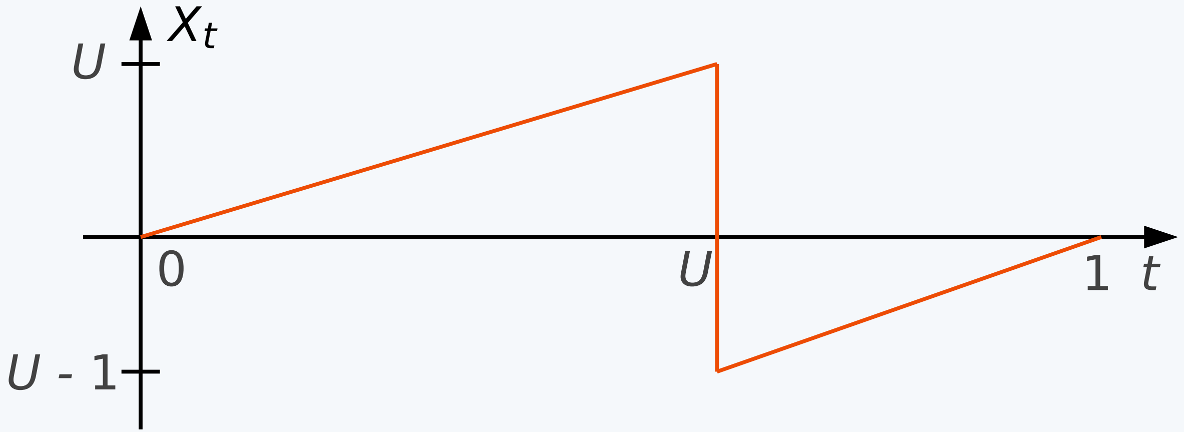


- Donoho, Vetterli, DeVore, and Daubechies ‘98 (“Ramp”); Meyer ‘92
- Image-like:

The Sawbridge

- Let U be unif[0,1] and

$$X(t) = t - \mathbf{1}(U \leq t) \quad t \in [0, 1]$$

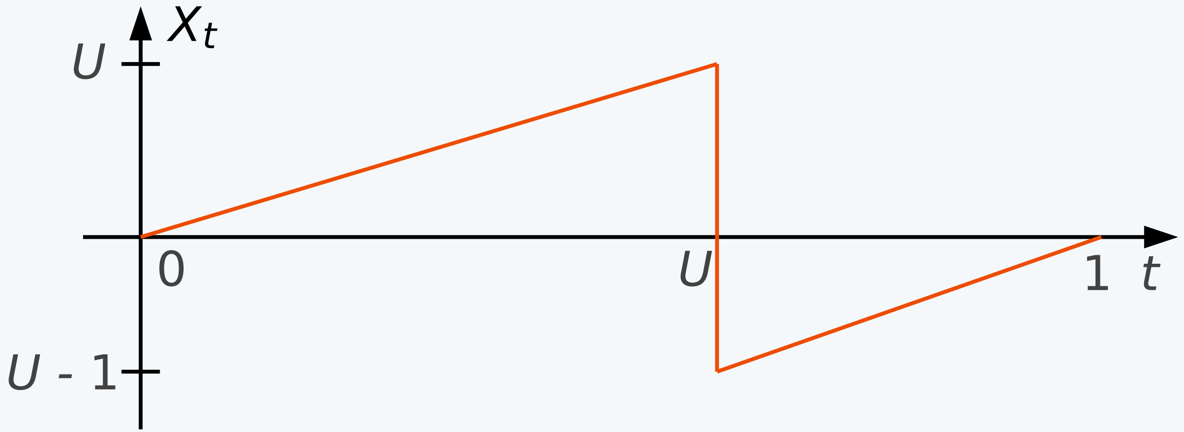


- Donoho, Vetterli, DeVore, and Daubechies '98 ("Ramp"); Meyer '92
- Image-like:
  - Two regions separated by a prominent edge (Donoho *et al.* '98)

The Sawbridge

- Let U be unif[0,1] and

$$X(t) = t - \mathbf{1}(U \leq t) \quad t \in [0,1]$$



- Donoho, Vetterli, DeVore, and Daubechies ‘98 (“Ramp”); Meyer ‘92
- Image-like:
  - Two regions separated by a prominent edge (Donoho *et al.* ‘98)
  - Support set is a 1-D manifold with infinite linear span

The KLT

Theorem (Wagner and Ballé): The sawbridge can be expanded as

$$X(t) = \sum_{k=1}^{\infty} Y_k \phi_k(t)$$

where $\{\phi_k(\cdot)\}_{k=1}^{\infty}$ is the orthonormal basis

$$\phi_k(t) = \sqrt{2} \cdot \sin(\pi k t)$$

and

$$Y_k = -\sqrt{2\lambda_k} \cos(\pi k U)$$

is a sequence of zero-mean, uncorrelated random variables, with Y_k having variance

$$\lambda_k = \frac{1}{\pi^2 k^2}.$$

The KLT

Theorem (Wagner and Ballé): The sawbridge can be expanded as

$$X(t) = \sum_{k=1}^{\infty} Y_k \phi_k(t)$$

Same autocorrelation as the Brownian Bridge

where $\{\phi_k(\cdot)\}_{k=1}^{\infty}$ is the orthonormal basis

$$\phi_k(t) = \sqrt{2} \cdot \sin(\pi k t)$$

and

$$Y_k = -\sqrt{2\lambda_k} \cos(\pi k U)$$

is a sequence of zero-mean, uncorrelated random variables, with Y_k having variance

$$\lambda_k = \frac{1}{\pi^2 k^2}.$$

The KLT

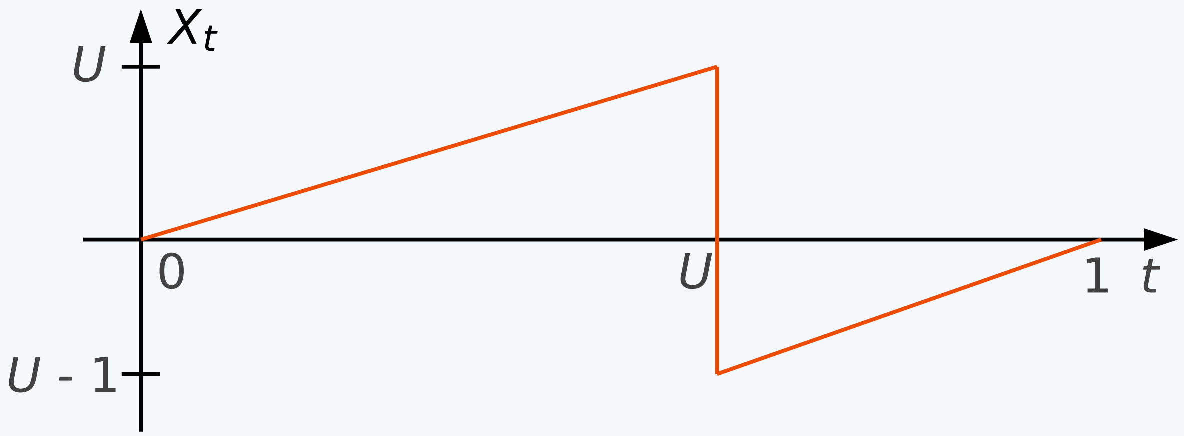
Corollary: If $f : L^2[0,1] \mapsto \mathbb{R}^k$ and $g : \mathbb{R}^k \mapsto L^2[0,1]$ then it is not possible to simultaneously satisfy the conditions:

1. f and g are linear
2. k is finite
3. $g(f(X(\cdot))) = X(\cdot) \quad \text{a.s.}$

The KLT

Corollary: If $f : L^2[0,1] \mapsto \mathbb{R}^k$ and $g : \mathbb{R}^k \mapsto L^2[0,1]$ then it is not possible to simultaneously satisfy the conditions:

1. f and g are linear
2. k is finite
3. $g(f(X(\cdot))) = X(\cdot) \quad$ a.s.



A Nonlinear Solution

- If

$$f(x(\cdot)) = \int_0^1 x(t)\,dt$$

$$g(y)(t) = t - \mathbf{1}(y + 1/2 \leq t)$$

A Nonlinear Solution

- If

$$f(x(\cdot)) = \int_0^1 x(t)\,dt$$

$$g(y)(t) = t - \mathbf{1}(y + 1/2 \leq t)$$

- Then

$$g(f(X(\cdot))) = X(\cdot) \quad \text{a.s.}$$

A Nonlinear Solution

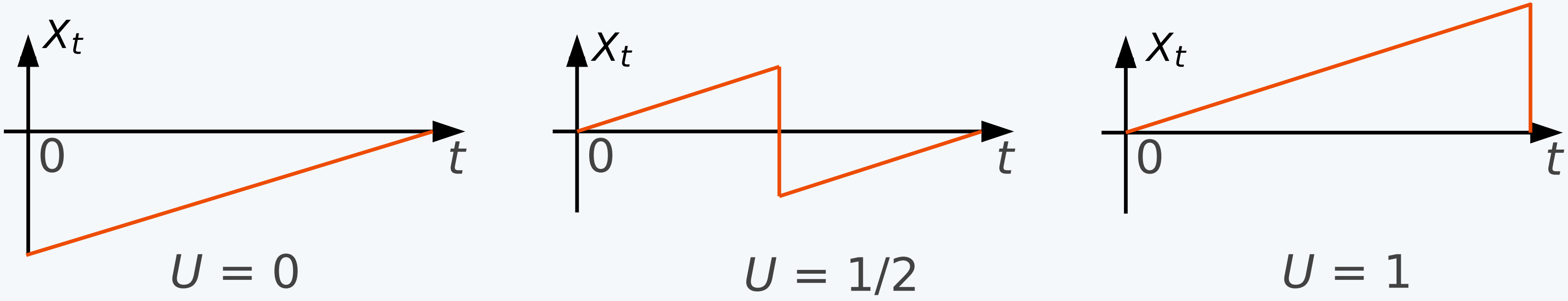
- If

$$f(x(\cdot)) = \int_0^1 x(t)\,dt$$

$$g(y)(t) = t - \mathbf{1}(y + 1/2 \leq t)$$

- Then

$$g(f(X(\cdot))) = X(\cdot) \quad \text{a.s.}$$



A Nonlinear Solution

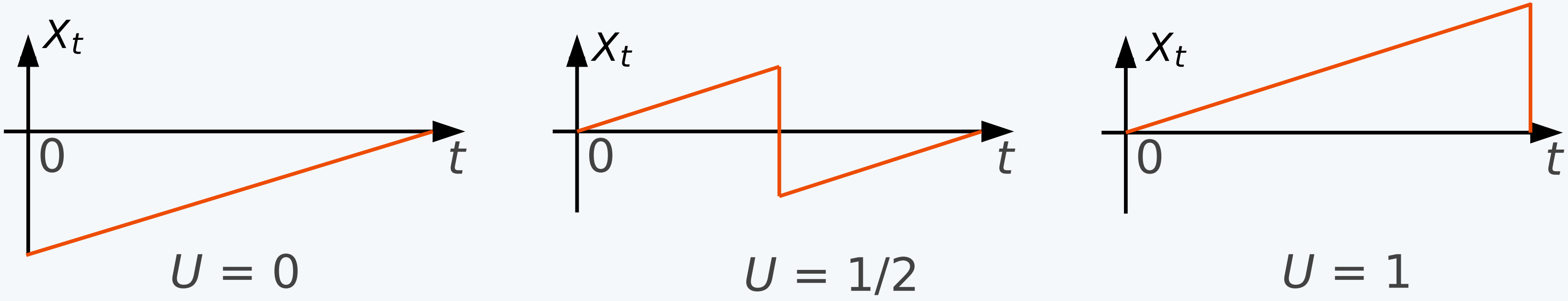
- If

$$f(x(\cdot)) = \int_0^1 x(t)\, dt$$

$$g(y)(t) = t - \mathbf{1}(y + 1/2 \leq t)$$

- Then

$$g(f(X(\cdot))) = X(\cdot) \quad \text{a.s.}$$



- Note that the "analysis" transform f is linear in this case.

Optimal Compression

Def: An *encoder* is a map $f : L^2[0,1] \mapsto \mathbb{N}$. The *entropy-distortion function* of the sawbridge is

$$H(\Delta) = \inf_f H(f(X(\cdot)))$$

$$\text{subject to } \Delta \geq E\left[\int_0^1 (X(t) - E[X(t)|f(X(\cdot))])^2 \, dt\right]$$

Optimal Compression

Theorem (Wagner and Ballé ’21): If $\Delta \geq 1/6$, then $H(\Delta) = 0$. For any $0 < \Delta < 1/6$, we have

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q,$$

where $q = \left(1 - \left\lfloor \frac{1}{p} \right\rfloor \cdot p\right)$ and p is the unique number in $(0, 1)$ such that

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta.$$

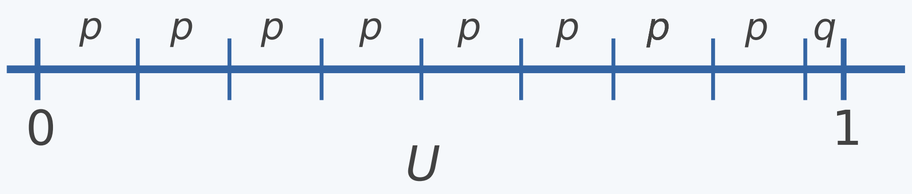
Optimal Compression

Theorem (Wagner and Ballé ’21): If $\Delta \geq 1/6$, then $H(\Delta) = 0$. For any $0 < \Delta < 1/6$, we have

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q,$$

where $q = \left(1 - \left\lfloor \frac{1}{p} \right\rfloor \cdot p\right)$ and p is the unique number in $(0,1)$ such that

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta.$$



Optimal Compression

Theorem (Wagner and Ballé '21): If $\Delta \geq 1/6$, then $H(\Delta) = 0$. For any $0 < \Delta < 1/6$, we have

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q,$$

where $q = \left(1 - \left\lfloor \frac{1}{p} \right\rfloor \cdot p\right)$ and p is the unique number in $(0, 1)$ such that

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta.$$

Optimal Compression

Theorem (Wagner and Ballé ’21): If $\Delta \geq 1/6$, then $H(\Delta) = 0$. For any $0 < \Delta < 1/6$, we have

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q,$$

where $q = \left(1 - \left\lfloor \frac{1}{p} \right\rfloor \cdot p\right)$ and p is the unique number in $(0, 1)$ such that

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta.$$

Proof steps:

Optimal Compression

Theorem (Wagner and Ballé '21): If $\Delta \geq 1/6$, then $H(\Delta) = 0$. For any $0 < \Delta < 1/6$, we have

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q,$$

where $q = \left(1 - \left\lfloor \frac{1}{p} \right\rfloor \cdot p\right)$ and p is the unique number in $(0, 1)$ such that

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta.$$

Proof steps:

- Any quantizer for $X(.)$ can be viewed as a quantizer for U.

Optimal Compression

Theorem (Wagner and Ballé '21): If $\Delta \geq 1/6$, then $H(\Delta) = 0$. For any $0 < \Delta < 1/6$, we have

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q,$$

where $q = \left(1 - \left\lfloor \frac{1}{p} \right\rfloor \cdot p\right)$ and p is the unique number in $(0, 1)$ such that

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta.$$

Proof steps:

- Any quantizer for $X(.)$ can be viewed as a quantizer for U.
- Show the best quantizers for U use contiguous cells.

Optimal Compression

Theorem (Wagner and Ballé '21): If $\Delta \geq 1/6$, then $H(\Delta) = 0$. For any $0 < \Delta < 1/6$, we have

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q,$$

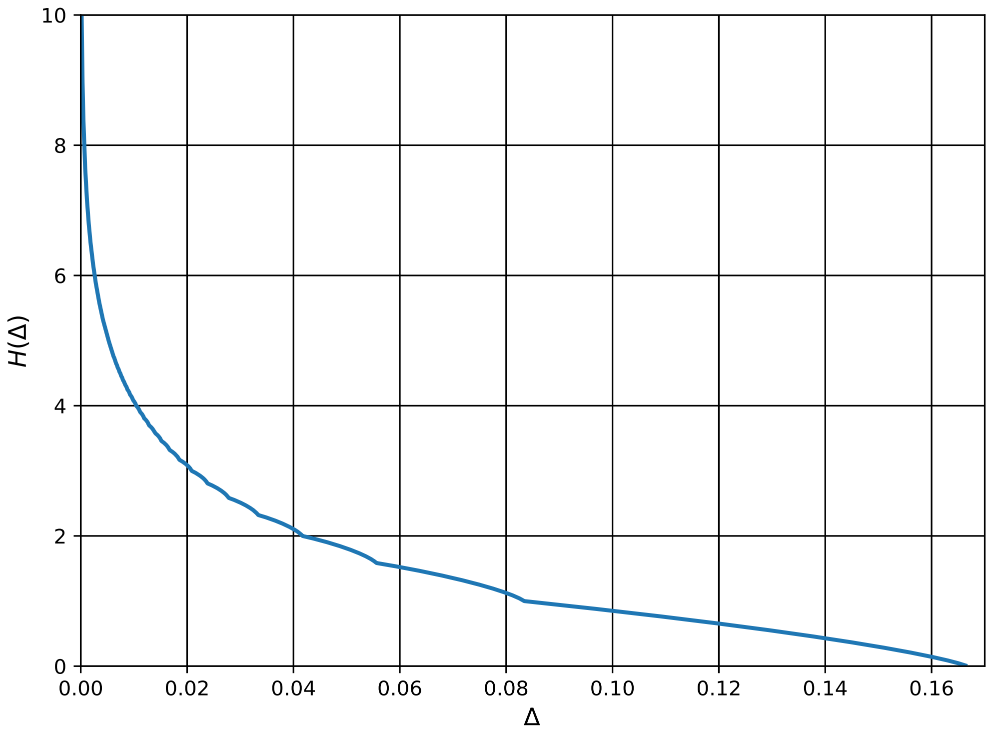
where $q = \left(1 - \left\lfloor \frac{1}{p} \right\rfloor \cdot p\right)$ and p is the unique number in $(0, 1)$ such that

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta.$$

Proof steps:

- Any quantizer for $X(.)$ can be viewed as a quantizer for U.
- Show the best quantizers for U use contiguous cells.
- Apply György and Linder '00 to solve nonconvex cell-size opt. problem

Optimal Compression



Optimal Compression

Theorem (Wagner and Ballé ’21): If $\Delta \geq 1/6$, then $H(\Delta) = 0$. For any $0 < \Delta < 1/6$, we have

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q,$$

where $q = \left(1 - \left\lfloor \frac{1}{p} \right\rfloor \cdot p\right)$ and p is the unique number in $(0, 1)$ such that

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta.$$

Optimal Compression

Theorem (Wagner and Ballé '21): If $\Delta \geq 1/6$, then $H(\Delta) = 0$. For any $0 < \Delta < 1/6$, we have

$$H(\Delta) = -\left\lfloor \frac{1}{p} \right\rfloor \cdot p \log p - q \log q,$$

where $q = \left(1 - \left\lfloor \frac{1}{p} \right\rfloor \cdot p\right)$ and p is the unique number in $(0, 1)$ such that

$$\left\lfloor \frac{1}{p} \right\rfloor \cdot p^2 + q^2 = 6\Delta.$$

Corollary (Wagner and Ballé '21): For the sawbridge,

$$\lim_{\Delta \to 0} \left| H(\Delta) - \log \frac{1}{6\Delta} \right| = 0.$$

KLT and Compression

Theorem: (Wagner and Ballé ’21): Let

$$H_{\text{KLT}}(\Delta) = \text{entropy of dithered quant.} + \text{waterfilling of KLT coefficients.}$$

Then

$$\lim_{\Delta \to 0} H_{\text{KLT}}(\Delta) \cdot \Delta = C$$

KLT and Compression

Theorem: (Wagner and Ballé ’21): Let

$$H_{\mathrm{KLT}}(\Delta) = \text{entropy of dithered quant.} + \text{waterfilling of KLT coefficients.}$$

Then

$$\begin{aligned}\lim_{\Delta\to 0} H_{\mathrm{KLT}}(\Delta)\cdot\Delta &= C \\ &\approx .3 \\ &= \frac{2}{\pi^2}\cdot\left(\int_0^1 h(s(\cdot)\star u_{\pi x\sqrt{12\gamma}}(\cdot))\,dx - \log(\pi\sqrt{12\gamma}/e)\right),\end{aligned}$$

where $h(\cdot)$ is differential entropy, $s(\cdot)$ is the arcsine density, $u_x(\cdot)$ is the uniform density over $[-x/2, x/2]$, γ is the unique solution to the fixed-point equation $\tan^{-1}(\pi\sqrt{\gamma}) = \frac{1}{\pi\sqrt{\gamma}}$ and $\star$ is convolution.

KLT and Compression

Theorem: (Wagner and Ballé '21): Let

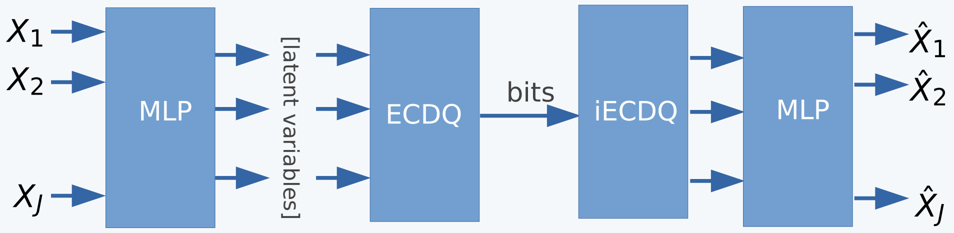
$$H_{\mathrm{KLT}}(\Delta) = \text{entropy of dithered quant.} + \text{waterfilling of KLT coefficients.}$$

Then

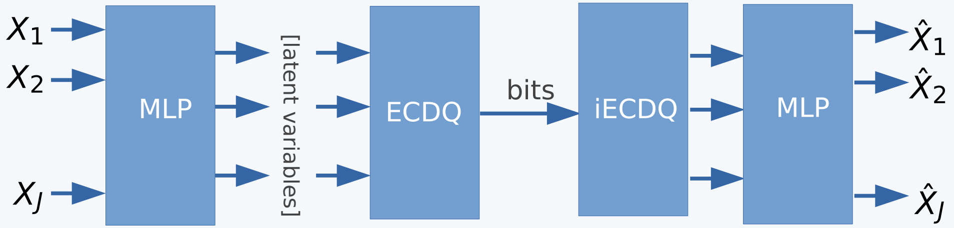
$$\lim_{\Delta \to 0} H_{\mathrm{KLT}}(\Delta) \cdot \Delta = C$$
$$\approx .3$$

So KLT + waterfilling + ECDQ is *exponentially* suboptimal
(cf. Donoho *et al.* '98)

Experimental Approach

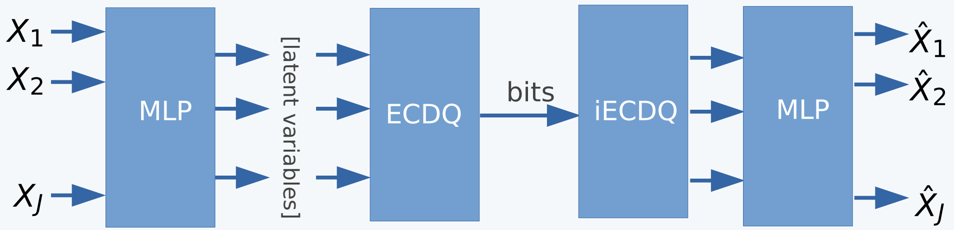


Experimental Approach



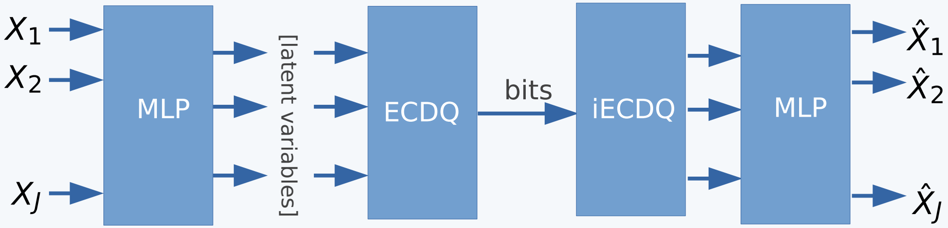
- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$

Experimental Approach



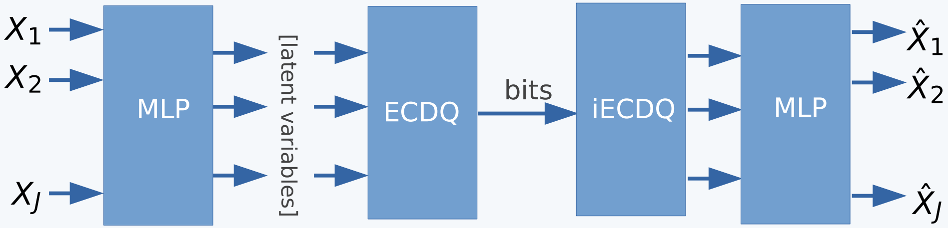
- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:

Experimental Approach



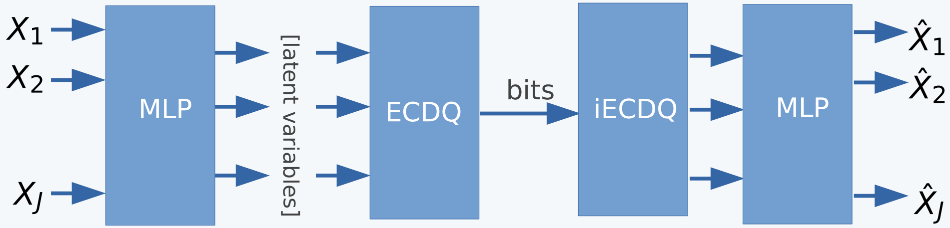
- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:
  - Have 3 layers with 100 nodes per layer (except last)

Experimental Approach



- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:
  - Have 3 layers with 100 nodes per layer (except last)
  - Have Leaky ReLU activation functions at each layer (except last)

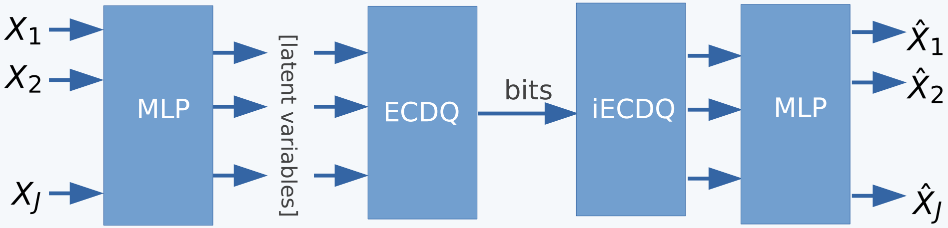
Experimental Approach



- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:
  - Have 3 layers with 100 nodes per layer (except last)
  - Have Leaky ReLU activation functions at each layer (except last)
- The overall system is trained via SGD to initially minimize the Lagrangian

$$\min_{f,g,q} E_{\vec{X},\vec{U}}\left[-\log q(f(\vec{X})+\vec{U})+\lambda||\vec{X}-g(f(\vec{X})+\vec{U})||^2\right]$$

Experimental Approach

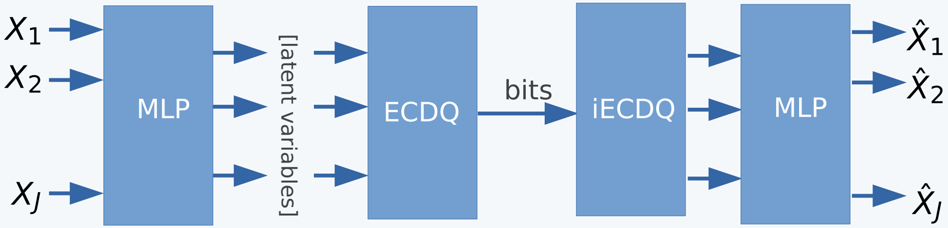


- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:
  - Have 3 layers with 100 nodes … cept last)
  - Have Leaky ReLU activation fu… ch layer (except last)
- The overall system is trained via SGD to initially minimize the Lagrangian

i.i.d. unif. dither

$$\min_{f,g,q} E_{\vec{X},\vec{U}}\left[-\log q(f(\vec{X})+\vec{U}) + \lambda ||\vec{X} - g(f(\vec{X})+\vec{U})||^2\right]$$

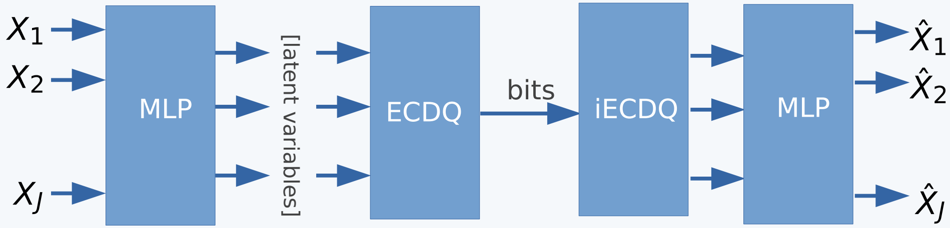
Experimental Approach



- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:
  - Have 3 layers with 100 nodes per layer (except last)
  - Have Leaky ReLU activation functions at each layer (except last)
- The overall system is trained via SGD to initially minimize the Lagrangian

$$\min_{f,g,q} E_{\vec{X},\vec{U}}\left[-\log q(f(\vec{X})+\vec{U}) + \lambda ||\vec{X} - g(f(\vec{X})+\vec{U})||^2\right]$$

Experimental Approach



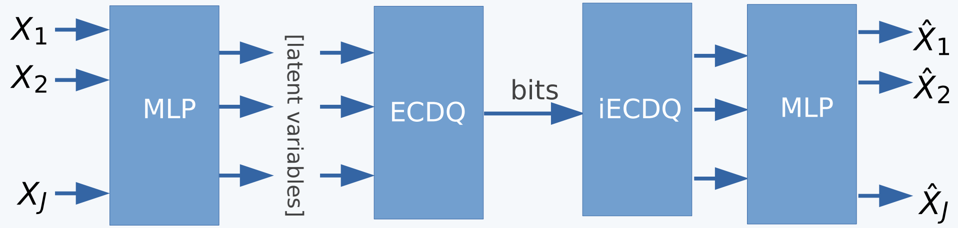
- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:
  - Have 3 layers with 100 nodes per layer (except last)
  - Have Leaky ReLU activation functions at each layer (except last)
- The overall system is trained via SGD to initially minimize the Lagrangian

$$\min_{f,g,q} E_{\vec{X},\vec{U}}\left[-\log q(f(\vec{X})+\vec{U}) + \lambda ||\vec{X} - g(f(\vec{X})+\vec{U})||^2\right]$$

which is gradually annealed to

$$\min_{f,g,q} E_{\vec{X}}\left[-\log q(Q(f(\vec{X}))) + \lambda ||\vec{X} - g(Q(f(\vec{X})))||^2\right]$$

Experimental Approach



- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:
  - Have 3 layers with 100 nodes per layer (except last)
  - Have Leaky ReLU activation functions at each layer (except last)
- The overall system is trained via SGD to initially minimize the Lagrangian

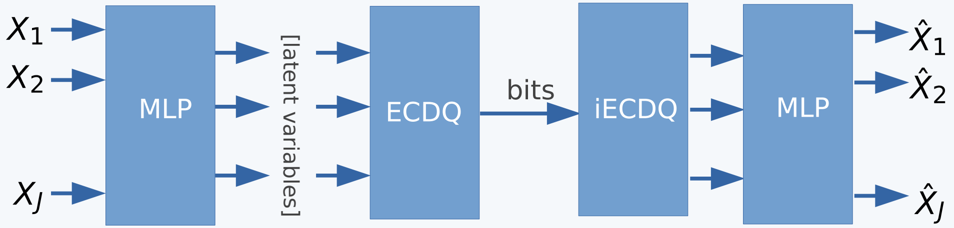
$$\min_{f,g,q} E_{\vec{X},\vec{U}}\left[-\log q(f(\vec{X}) + \vec{U}) + \lambda ||\vec{X} - \ldots ||^2\right]$$

hard quantizer

which is gradually annealed to

$$\min_{f,g,q} E_{\vec{X}}\left[-\log q(Q(f(\vec{X}))) + \lambda ||\vec{X} - g(Q(f(\vec{X})))||^2\right]$$

Experimental Approach



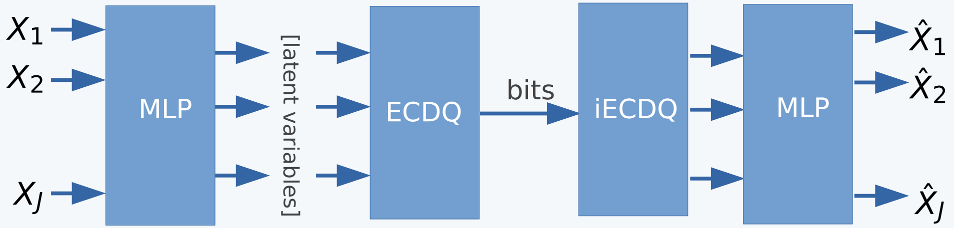
- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:
  - Have 3 layers with 100 nodes per layer (except last)
  - Have Leaky ReLU activation functions at each layer (except last)
- The overall system is trained via SGD to initially minimize the Lagrangian

$$\min_{f,g,q} E_{\vec{X},\vec{U}}\left[-\log q(f(\vec{X})+\vec{U})+\lambda||\vec{X}-g(f(\vec{X})+\vec{U})||^2\right]$$

which is gradually annealed to

$$\min_{f,g,q} E_{\vec{X}}\left[-\log q(Q(f(\vec{X})))+\lambda||\vec{X}-g(Q(f(\vec{X})))||^2\right]$$

Experimental Approach



- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:
  - Have 3 layers with 100 nodes per layer (except last)
  - Have Leaky ReLU activation functions at each layer (except last)
- The overall system is trained via SGD to initially minimize the Lagrangian

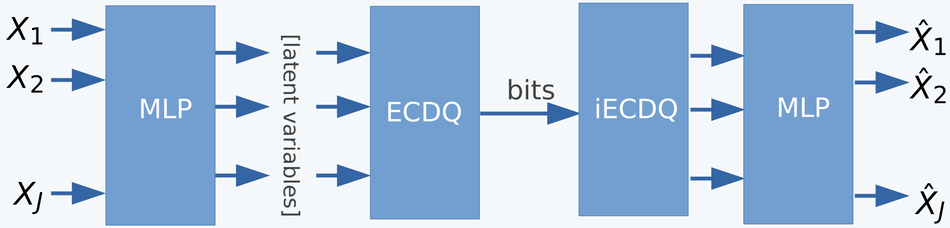
$$\min_{f,g,q} E_{\vec{X},\vec{U}}\left[-\log q(f(\vec{X})+\vec{U}) + \lambda||\vec{X} - g(f(\vec{X})+\vec{U})||^2\right]$$

which is gradually annealed to

$$\min_{f,g,q} E_{\vec{X}}\left[-\log q(Q(f(\vec{X}))) + \lambda||\vec{X} - g(Q(f(\vec{X})))||^2\right]$$

[Agustsson and Theis ‘20]

Experimental Approach



- Sample sawbridge at $J = 1024$ points to create vector $\vec{X}$
- The MLPs:
  - Have 3 layers with 100 nodes per layer (except last)
  - Have Leaky ReLU activation functions at each layer (except last)
- The overall system is trained via SGD to initially minimize the Lagrangian

$$\min_{f,g,q} E_{\vec{X},\vec{U}}\left[-\log q(f(\vec{X})+\vec{U})+\lambda||\vec{X}-g(f(\vec{X})+\vec{U})||^2\right]$$

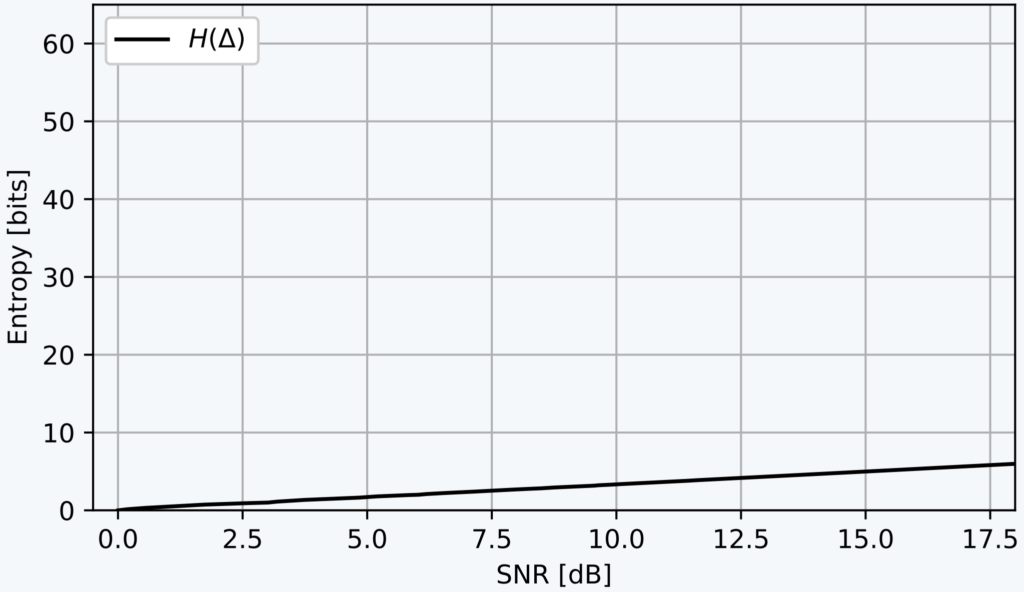
which is gradually annealed to

$$\min_{f,g,q} E_{\vec{X}}\left[-\log q(Q(f(\vec{X})))+\lambda||\vec{X}-g(Q(f(\vec{X})))||^2\right]$$

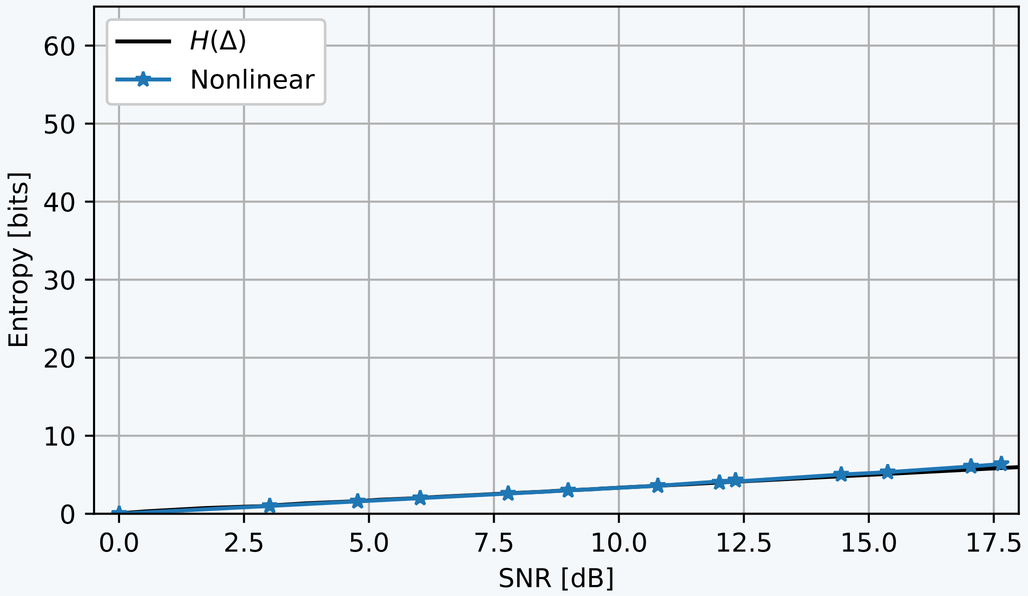
[Agustsson and Theis '20]

- For linear transforms, we take single-layer MLPs with affine activations

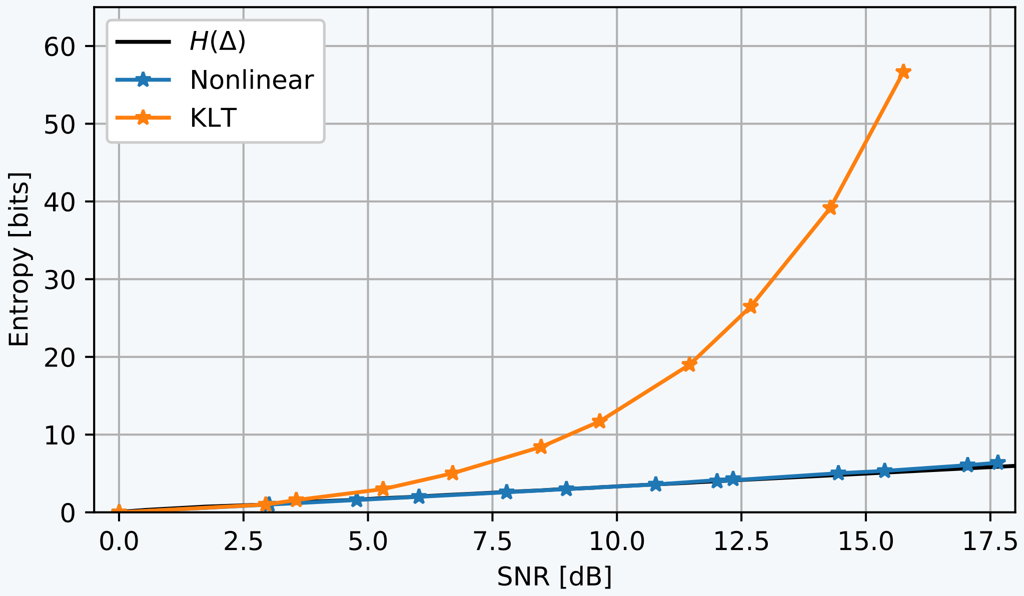
Numerical Results



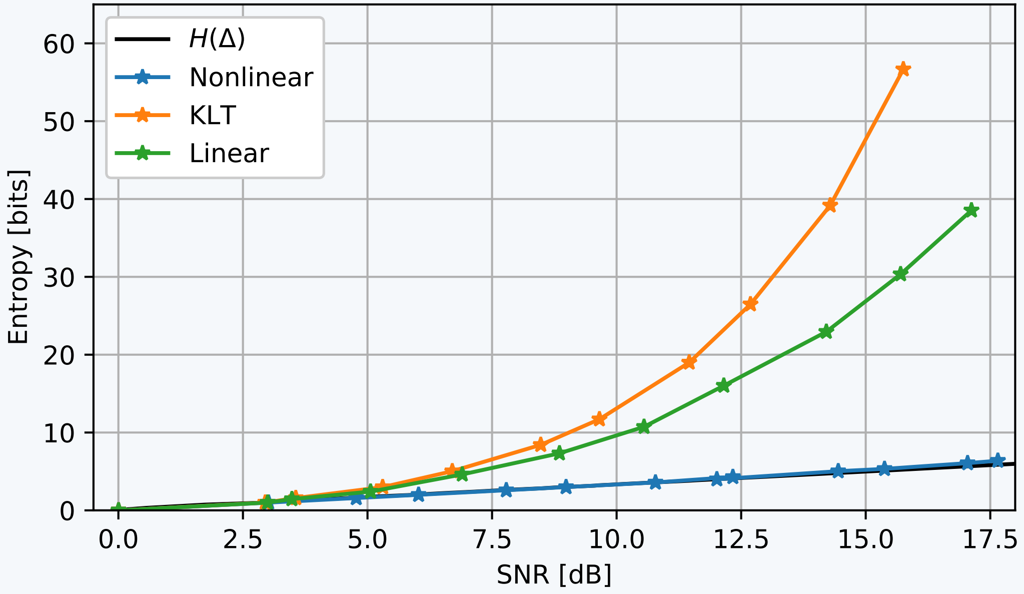
Numerical Results



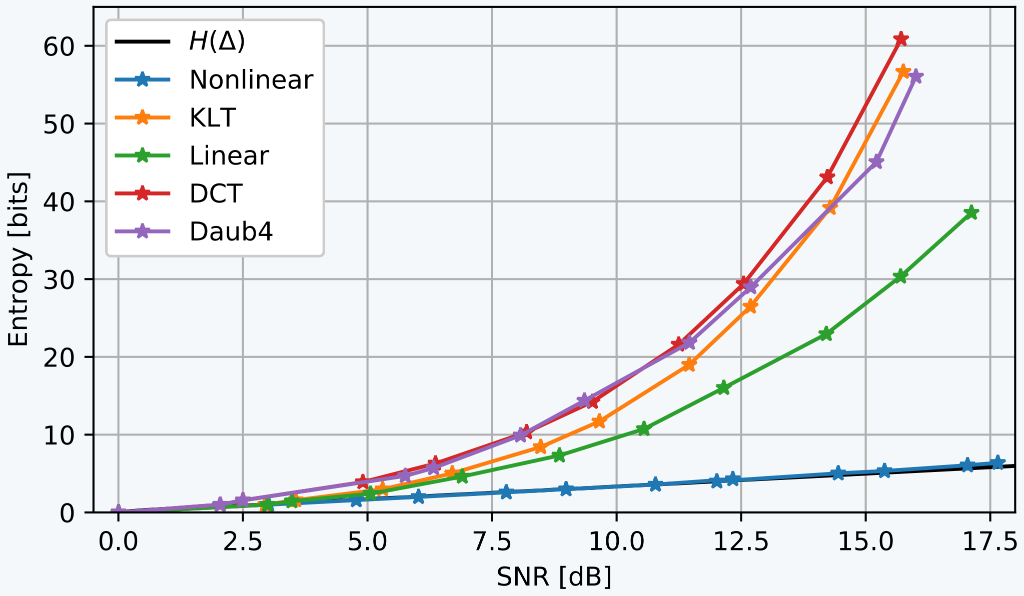
Numerical Results



Numerical Results



Numerical Results



KLT and Compression

Theorem: (Wagner and Ballé '21): Let

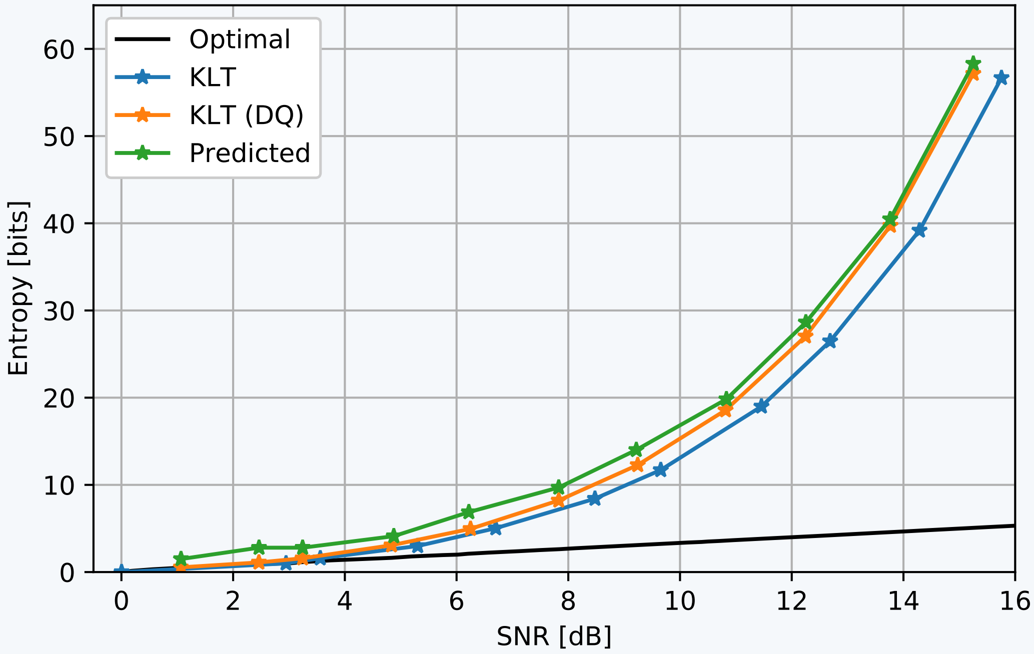
$$H_{\text{KLT}}(\Delta) = \text{entropy of dithered quant.} + \text{waterfilling of KLT coefficients.}$$

Then

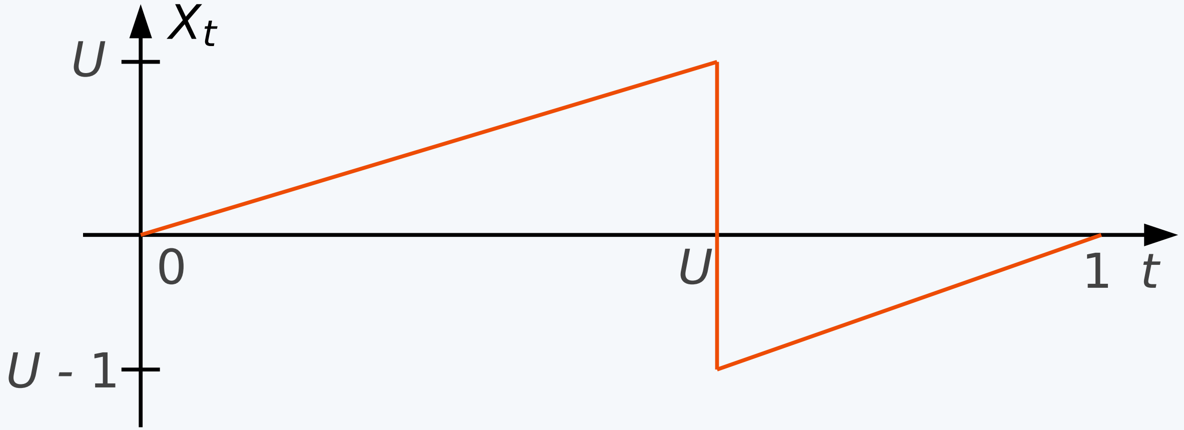
$$\begin{aligned}\lim_{\Delta \to 0} H_{\text{KLT}}(\Delta) \cdot \Delta &= C \\ &\approx .3 \\ &= \frac{2}{\pi^2} \cdot \left(\int_0^1 h(s(\cdot) \star u_{\pi x \sqrt{12\gamma}}(\cdot))\, dx - \log(\pi \sqrt{12\gamma}/e) \right),\end{aligned}$$

where $h(\cdot)$ is differential entropy, $s(\cdot)$ is the arcsine density, $u_x(\cdot)$ is the uniform density over $[-x/2, x/2]$, γ is the unique solution to the fixed-point equation $\tan^{-1}(\pi\sqrt{\gamma}) = \frac{1}{\pi\sqrt{\gamma}}$ and $\star$ is convolution.

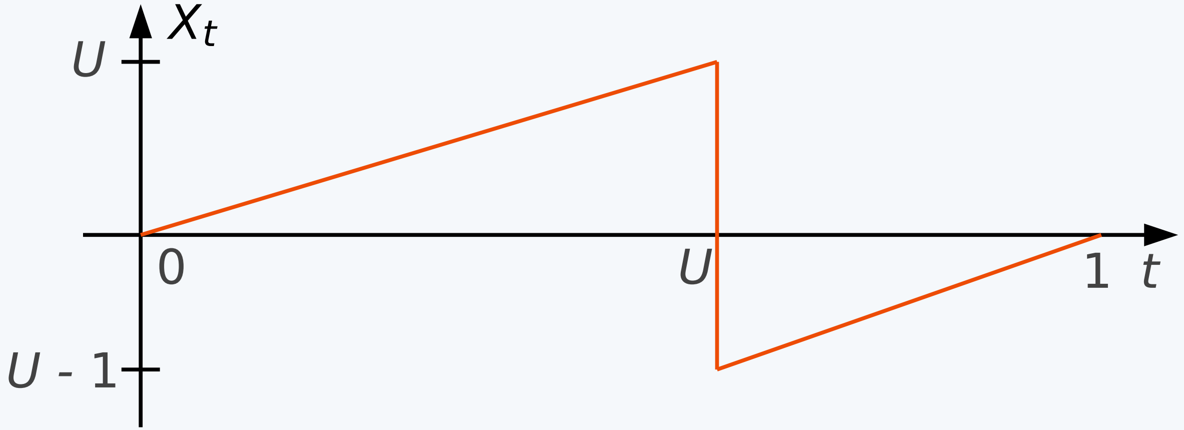
Numerical Results



Conclusion

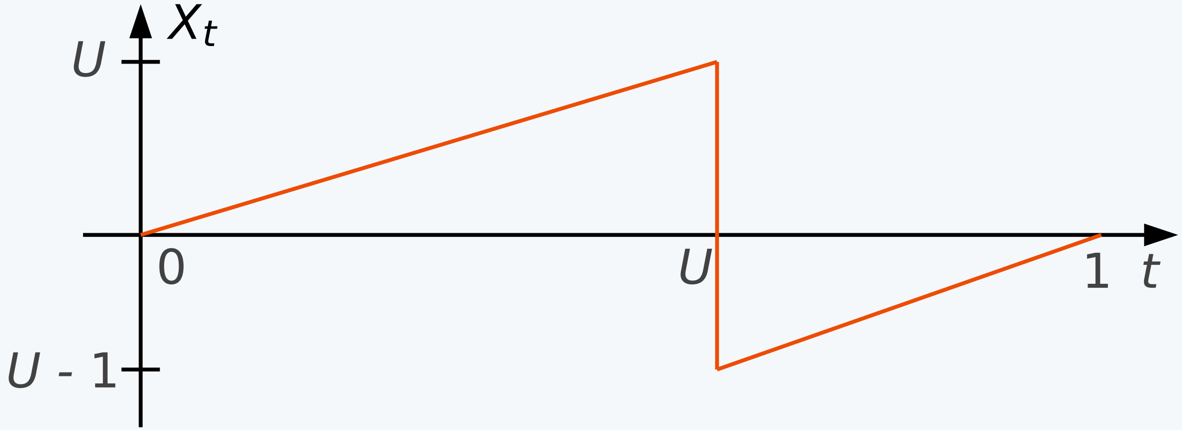


Conclusion



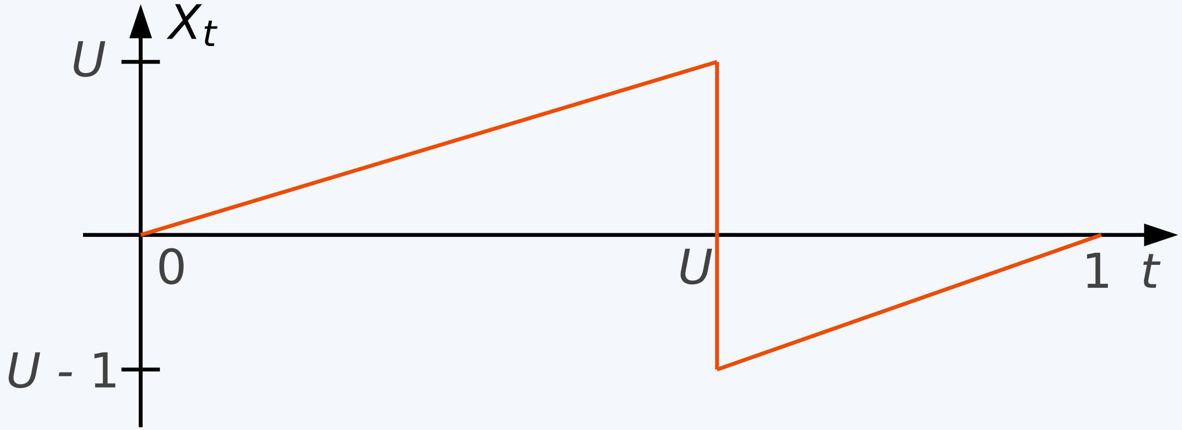
- The sawbridge is a simple image model capturing edges and manifold structure for which:

Conclusion



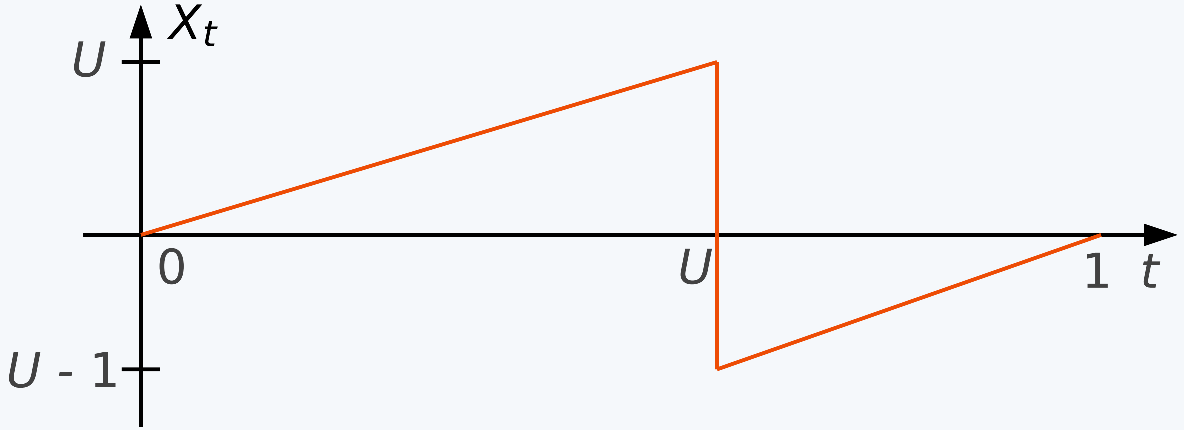
- The sawbridge is a simple image model capturing edges and manifold structure for which:
  - The optimal compressor can be exactly characterized.

Conclusion



- The sawbridge is a simple image model capturing edges and manifold structure for which:
  - The optimal compressor can be exactly characterized.
  - Trained ANNs are (numerically) optimal and beat the classical KLT-based approach by an exponential margin.

Conclusion



- The sawbridge is a simple image model capturing edges and manifold structure for which:
  - The optimal compressor can be exactly characterized.
  - Trained ANNs are (numerically) optimal and beat the classical KLT-based approach by an exponential margin.
- Provides one answer to the question “For what sources are artificial neural networks optimal compressors?”

Conclusion

Extended version:

https://arxiv.org/abs/2011.05065

Code:

https://github.com/tensorflow/compression/tree/master/models/toy_sources