Information-Theoretic Modeling

Lecture 8: Universal Source Coding

Jyrki Kivinen

Department of Computer Science, University of Helsinki

Autumn 2012
Lecture 8: Universal Source Coding


Jyrki Kivinen
Information-Theoretic Modeling
1 Universal Source Codes
   • Definitions
   • Universal Models
1. Universal Source Codes
   - Definitions
   - Universal Models

2. Two-Part Codes
   - Discrete Parameters
   - Continuous Parameters
   - Asymptotics: $\frac{k}{2} \log n$
1 Universal Source Codes
   - Definitions
   - Universal Models

2 Two-Part Codes
   - Discrete Parameters
   - Continuous Parameters
   - Asymptotics: $\frac{k}{2} \log n$

3 Advanced Universal Codes
   - Mixture Codes
   - Normalized Maximum Likelihood
   - Universal Prediction
Definitions

Our basic setting is that we have some data \( D = (x_1, \ldots, x_m) \) where the individual data points \( x_i \) come from some domain \( \mathcal{X} \).
Our basic setting is that we have some data \( D = (x_1, \ldots, x_m) \) where the individual data points \( x_i \) come from some domain \( \mathcal{X} \).

We write \( \mathcal{D} \) for the set of all possible data. A typical situation is \( \mathcal{D} = \mathcal{X}^n \) where \( n \) may or may not be known in advance.
Our basic setting is that we have some data \( D = (x_1, \ldots, x_m) \) where the individual data points \( x_i \) come from some domain \( \mathcal{X} \).

We write \( D \) for the set of all possible data. A typical situation is \( D = \mathcal{X}^n \) where \( n \) may or may not be known in advance.

A probability distribution \( p \) over \( D \) is called a model.
Our basic setting is that we have some data $D = (x_1, \ldots, x_m)$ where the individual data points $x_i$ come from some domain $\mathcal{X}$.

We write $\mathcal{D}$ for the set of all possible data. A typical situation is $\mathcal{D} = \mathcal{X}^n$ where $n$ may or may not be known in advance.

A probability distribution $p$ over $\mathcal{D}$ is called a model.

A set of models $\mathcal{M}$ is called a model class.
Our basic setting is that we have some data $D = (x_1, \ldots, x_m)$ where the individual data points $x_i$ come from some domain $\mathcal{X}$.

We write $\mathcal{D}$ for the set of all possible data. A typical situation is $\mathcal{D} = \mathcal{X}^n$ where $n$ may or may not be known in advance.

A probability distribution $p$ over $\mathcal{D}$ is called a model.

A set of models $\mathcal{M}$ is called a model class.

Model classes are often parametric: $\mathcal{M} = \{ p_\theta \mid \theta \in \Theta \}$ where $\Theta \subseteq \mathbb{R}^k$ for some $k$ and $p_\theta$ is a model for each $\theta \in \Theta$. 
Example Let $p_{\mu,\sigma^2}$ be the normal distribution over $\mathcal{X} = \mathbb{R}$ with mean $\mu$ and variance $\sigma^2$. 
Example Let $p_{\mu,\sigma^2}$ be the normal distribution over $\mathcal{X} = \mathbb{R}$ with mean $\mu$ and variance $\sigma^2$.

We have a parametric family $\mathcal{M} = \{ p_{\theta} \mid \theta \in \Theta \}$ where

$\Theta = \{ (\mu, \sigma^2) \in \mathbb{R}^2 \mid \sigma^2 > 0 \}$.
**Example** Let $p_{\mu,\sigma^2}$ be the normal distribution over $\mathcal{X} = \mathbb{R}$ with mean $\mu$ and variance $\sigma^2$.

We have a parametric family $\mathcal{M} = \{ p_\theta \mid \theta \in \Theta \}$ where $\Theta = \{ (\mu, \sigma^2) \in \mathbb{R}^2 \mid \sigma^2 > 0 \}$.

We can extend $p_{\mu,\sigma^2}$ into a distribution $p^{(n)}_{\mu,\sigma^2}$ over $\mathcal{D} = \mathbb{R}^n$ by assuming independence: $p^{(n)}_{\mu,\sigma^2}(x_1, \ldots, x_n) = p_{\mu,\sigma^2}(x_1) \cdots p_{\mu,\sigma^2}(x_n)$. 
Example Let $p_{\mu,\sigma^2}$ be the normal distribution over $\mathcal{X} = \mathbb{R}$ with mean $\mu$ and variance $\sigma^2$.

We have a parametric family $\mathcal{M} = \{ p_\theta \mid \theta \in \Theta \}$ where

$$\Theta = \{ (\mu, \sigma^2) \in \mathbb{R}^2 \mid \sigma^2 > 0 \}.$$

We can extend $p_{\mu,\sigma^2}$ into a distribution $p^{(n)}_{\mu,\sigma^2}$ over $\mathcal{D} = \mathbb{R}^n$ by assuming independence:

$$p^{(n)}_{\mu,\sigma^2}(x_1, \ldots, x_n) = p_{\mu,\sigma^2}(x_1) \cdots p_{\mu,\sigma^2}(x_n).$$

We often abuse notation by just writing $p_\theta(x_1, \ldots, x_n)$ instead of $p^{(n)}_\theta(x_1, \ldots, x_n)$. 
**Example** Let $p_{\mu,\sigma^2}$ be the normal distribution over $\mathcal{X} = \mathbb{R}$ with mean $\mu$ and variance $\sigma^2$. We have a parametric family $\mathcal{M} = \{ p_{\theta} \mid \theta \in \Theta \}$ where $\Theta = \{ (\mu, \sigma^2) \in \mathbb{R}^2 \mid \sigma^2 > 0 \}$.

We can extend $p_{\mu,\sigma^2}$ into a distribution $p_{\mu,\sigma^2}^{(n)}$ over $\mathcal{D} = \mathbb{R}^n$ by assuming independence: $p_{\mu,\sigma^2}^{(n)}(x_1, \ldots, x_n) = p_{\mu,\sigma^2}(x_1) \ldots p_{\mu,\sigma^2}(x_n)$.

We often abuse notation by just writing $p_{\theta}(x_1, \ldots, x_n)$ instead of $p_{\theta}^{(n)}(x_1, \ldots, x_n)$.

However, keep in mind that we may also have $p$ over $\mathcal{D}$ that does not satisfy the independence assumption.
Information-theoretic modeling?

In what follows, it’s important to keep in mind that we don’t claim that we can find a “true” model $p$ that “really” generated the data $D$, or even that such a “true” model exists.
Information-theoretic modeling?

In what follows, it’s important to keep in mind that we don’t claim that we can find a “true” model $p$ that “really” generated the data $D$, or even that such a “true” model exists.

However, keeping in mind how codes and distributions are related, it seems reasonable to think that
In what follows, it’s important to keep in mind that we don’t claim that we can find a “true” model $p$ that “really” generated the data $D$, or even that such a “true” model exists.

However, keeping in mind how codes and distributions are related, it seems reasonable to think that

If a code based on model $p$ is good at compressing $D$, then perhaps studying $p$ can tell us something useful about $D$. 
The model within $\mathcal{M}$ that achieves the shortest code-length for data $x$ is the **maximum likelihood (ML) model**:

$$
\min_{\theta \in \Theta} \log_2 \frac{1}{p_\theta(D)} = \log_2 \frac{1}{p_{\hat{\theta}}(D)}.
$$
The model within $\mathcal{M}$ that achieves the shortest code-length for data $x$ is the \textbf{maximum likelihood (ML) model}:

$$\min_{\theta \in \Theta} \log_2 \frac{1}{p_{\theta}(D)} = \log_2 \frac{1}{p_{\hat{\theta}}(D)}.$$  

\textbf{Depends on $D$!}
Definitions

The model within $\mathcal{M}$ that achieves the shortest code-length for data $x$ is the **maximum likelihood (ML) model**:

$$\min_{\theta \in \Theta} \log_2 \frac{1}{p_{\theta}(D)} = \log_2 \frac{1}{p_{\hat{\theta}}(D)}.$$  

[Depends on $D$!]

For model $q$, the excess code-length or “regret” over the ML model in $\mathcal{M}$ is given by

$$\log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)}.$$
A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a universal model (code) relative to model class $\mathcal{M}$:

$$\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\theta(D)} \right] = 0 .$$

(1)
Universal models

**Universal model**

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a **universal model** (code) relative to model class $\mathcal{M}$:

$$
\lim_{n \to \infty} \frac{1}{n} \max_{D \in D} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \right] = 0 \quad (1)
$$

\[ \log_2 \frac{1}{p_\hat{\theta}(D)} \leq \log_2 \frac{1}{p_\theta(D)} \]
Universal models

Universal model

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a **universal model** (code) relative to model class $\mathcal{M}$:

$$
\lim_{{n \to \infty}} \frac{1}{n} \max_{{D \in \mathcal{D}}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \right] = 0 .
$$

$$
- \log_2 \frac{1}{p_\hat{\theta}(D)} \geq - \log_2 \frac{1}{p_\theta(D)}
$$
Universal models

Universal model

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a **universal model** (code) relative to model class $\mathcal{M}$:

$$\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \right] = 0 .$$

$$\log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \geq \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\theta(D)}$$
Universal models

Universal model

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a universal model (code) relative to model class $\mathcal{M}$:

$$\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \right] = 0 . \quad (1)$$

$$E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \right] 
\geq E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\theta(D)} \right]$$
Universal models

Universal model

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a **universal model** (code) relative to model class $\mathcal{M}$:

$$\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)} \right] = 0 . \quad (1)$$

$$E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)} \right] \geq E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} \right] - E_{D \sim p_\theta} \left[ \log_2 \frac{1}{p_\theta(D)} \right]$$

This is another (stochastic) definition of universality, equivalent to $1/n \rightarrow 0$ for all $\theta \in \Theta$. It is weaker since (1) $\Rightarrow$ (2).
A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a **universal model** (code) relative to model class $\mathcal{M}$:

$$
\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)} \right] = 0 . \quad (1)
$$

$$
E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)} \right] \\
\geq E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} \right] - \sum_D p_\theta(D) \log_2 \frac{1}{p_\theta(D)}
$$
Universal models

Universal model

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a universal model (code) relative to model class $\mathcal{M}$:

$$\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \right] = 0 \ .$$

(1)

$$E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \right] \geq E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} \right] - H(p_\theta)$$
Universal models

**Universal model**

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a **universal model** (code) relative to model class $\mathcal{M}$:

$$
\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \right] = 0 .
$$

$$
E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \right] \geq E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} \right] - nH(p_\theta^{(1)})
$$
Universal models

Universal model

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a universal model (code) relative to model class $\mathcal{M}$:

$$\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\theta(D)} \right] = 0 .$$  \hspace{1cm} (1)

$$\frac{1}{n} \mathbb{E}_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\theta(D)} \right] \geq \frac{1}{n} \mathbb{E}_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} \right] - H(p_\theta^{(1)})$$
Universal models

Universal model

A model (code) whose regret grows slower than \( n \), for all data sequences, is said to be a \textbf{universal model} (code) relative to model class \( \mathcal{M} \):

\[
\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)} \right] = 0 . \tag{1}
\]

\[
\lim_{n \to \infty} \frac{1}{n} E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)} \right] \geq \lim_{n \to \infty} \frac{1}{n} E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} \right] - H(p_{\theta}^{(1)})
\]
Universal models

Universal model

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a **universal model** (code) relative to model class $\mathcal{M}$:

\[
\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\theta(D)} \right] = 0.
\]

This is another (stochastic) definition of universality, equivalent to $1/n D_\theta(D_q) \to 0$ for all $\theta \in \Theta$. It is weaker since $1 \Rightarrow 2$.

\[
0 \geq \lim_{n \to \infty} \frac{1}{n} \mathbb{E}_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} \right] - H(p_\theta^{(1)})
\]
Universal models

A model (code) whose regret grows slower than \( n \), for all data sequences, is said to be a \textbf{universal model} (code) relative to model class \( \mathcal{M} \):

\[
\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)} \right] = 0 .
\]  \quad (1)

\[
\lim_{n \to \infty} \frac{1}{n} E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} \right] \leq H(p_\theta^{(1)})
\]
Universal models

Universal model

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a **universal model** (code) relative to model class $\mathcal{M}$:

$$\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_\hat{\theta}(D)} \right] = 0 \ .$$  \hspace{1cm} (1)

$$\lim_{n \to \infty} \frac{1}{n} \mathbb{E}_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} \right] = H(p_\theta^{(1)}) \quad \text{ (2)}$$

This is another (stochastic) definition of universality, equivalent to $1/n D(\|q\|_\theta) \to 0$ for all $\theta \in \Theta$. It is weaker since (1) $\Rightarrow$ (2).
Universal models

Universal model

A model (code) whose regret grows slower than $n$, for all data sequences, is said to be a \textbf{universal model} (code) relative to model class $\mathcal{M}$:

$$
\lim_{n \to \infty} \frac{1}{n} \max_{D \in \mathcal{D}} \left[ \log_2 \frac{1}{q(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)} \right] = 0 \ . 
$$

This is another (stochastic) definition of universality, equivalent to $\frac{1}{n} D(p_\theta \parallel q) \to 0$ for all $\theta \in \Theta$. It is weaker since (1) $\Rightarrow$ (2).

$$
\lim_{n \to \infty} \frac{1}{n} E_{D \sim p_\theta} \left[ \log_2 \frac{1}{q(D)} \right] = H(p_\theta^{(1)}) 
$$
The typical situation might be as follows:
Universal models

The typical situation might be as follows:

1. We know (think) that the source symbols are generated by a Bernoulli model with parameter $p \in [0, 1]$. 
Universal models

The typical situation might be as follows:

1. We know (think) that the source symbols are generated by a Bernoulli model with parameter $p \in [0, 1]$.
2. We’d like to encode data at rate $H(p)$. 
The typical situation might be as follows:

1. We know (think) that the source symbols are generated by a Bernoulli model with parameter $p \in [0, 1]$.
2. We’d like to encode data at rate $H(p)$.
3. However, we do not know $p$ in advance.
The typical situation might be as follows:

1. We know (think) that the source symbols are generated by a Bernoulli model with parameter $p \in [0, 1]$.
2. We’d like to encode data at rate $H(p)$.
3. However, we do not know $p$ in advance.

Again, we don’t need to believe that data are really generated by a Bernoulli model.
1 Universal Source Codes
   • Definitions
   • Universal Models

2 Two-Part Codes
   • Discrete Parameters
   • Continuous Parameters
   • Asymptotics: $\frac{k}{2} \log n$

3 Advanced Universal Codes
   • Mixture Codes
   • Normalized Maximum Likelihood
   • Universal Prediction
Let $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ be a parametric probabilistic model class.
Two-Part Codes

Let $\mathcal{M} = \{ p_\theta : \theta \in \Theta \}$ be a parametric probabilistic model class. If the parameter space $\Theta$ is discrete, we can construct a (prefix) code $C_1 : \Theta \rightarrow \{0,1\}^*$ which maps each parameter value to a codeword of length $\ell_1(\theta)$. 
Two-Part Codes

Let \( \mathcal{M} = \{ p_\theta : \theta \in \Theta \} \) be a parametric probabilistic model class.

If the parameter space \( \Theta \) is discrete, we can construct a (prefix) code \( C_1 : \Theta \rightarrow \{0, 1\}^* \) which maps each parameter value to a codeword of length \( \ell_1(\theta) \).

For any distribution \( p_\theta \), the Shannon code-lengths satisfy

\[
\ell_\theta(D) = \left\lceil \log_2 \frac{1}{p_\theta(D)} \right\rceil \approx \log_2 \frac{1}{p_\theta(D)} .
\]
Two-Part Codes

Let $\mathcal{M} = \{p_\theta : \theta \in \Theta\}$ be a parametric probabilistic model class.

If the parameter space $\Theta$ is discrete, we can construct a (prefix) code $C_1 : \Theta \to \{0, 1\}^*$ which maps each parameter value to a codeword of length $\ell_1(\theta)$.

For any distribution $p_\theta$, the Shannon code-lengths satisfy

$$\ell_\theta(D) = \left\lceil \log_2 \frac{1}{p_\theta(D)} \right\rceil \approx \log_2 \frac{1}{p_\theta(D)}.$$

Using parameter value $\theta$, the total code-length becomes ($\approx$)

$$\ell_1(\theta) + \log_2 \frac{1}{p_\theta(D)}.$$
Two-Part Codes

Using the maximum likelihood parameter, the total code-length becomes

\[ \ell_{\text{two-part}}(D) = \ell_1(\hat{\theta}) + \log_2 \frac{1}{p_{\hat{\theta}}(D)}. \]
Two-Part Codes

Using the maximum likelihood parameter, the total code-length becomes

$$\ell_{\text{two-part}}(D) = \ell_1(\hat{\theta}) + \log_2 \frac{1}{p_{\hat{\theta}}(D)}.$$  

Hence, the regret of the two-part code is

$$\ell_{\text{two-part}}(D) - \log_2 \frac{1}{p_{\hat{\theta}}(D)} = \ell_1(\hat{\theta}).$$
Two-Part Codes

Using the maximum likelihood parameter, the total code-length becomes

\[ \ell_{\text{two-part}}(D) = \ell_1(\hat{\theta}) + \log_2 \left( \frac{1}{p_{\hat{\theta}}(D)} \right) . \]

Hence, the regret of the two-part code is

\[ \ell_{\text{two-part}}(D) - \log_2 \left( \frac{1}{p_{\hat{\theta}}(D)} \right) = \ell_1(\hat{\theta}) < cn \quad \text{for all } c > 0 \text{ and large } n. \]
Two-Part Codes

Using the maximum likelihood parameter, the total code-length becomes

$$\ell_{\text{two-part}}(D) = \ell_1(\hat{\theta}) + \log_2 \frac{1}{p_{\hat{\theta}}(D)}.$$ 

Hence, the regret of the two-part code is

$$\ell_{\text{two-part}}(D) - \log_2 \frac{1}{p_{\hat{\theta}}(D)} = \ell_1(\hat{\theta}) < cn \quad \text{for all } c > 0 \text{ and large } n.$$ 

For discrete parameter models the two-part code is universal.
Universality of Two-Part Codes

However, keep in mind that universality is not everything.
Universality of Two-Part Codes

However, keep in mind that universality is not everything.

Since the two-part code is universal, its regret goes to zero, but there may be other codes for which regret goes to zero faster.
Universality of Two-Part Codes

However, keep in mind that universality is not everything.

Since the two-part code is universal, its regret goes to zero, but there may be other codes for which regret goes to zero faster.

On the other hand, two-part codes have the advantage of being reasonably easy to understand.
Universality of Two-Part Codes

However, keep in mind that universality is not everything.

Since the two-part code is universal, its regret goes to zero, but there may be other codes for which regret goes to zero faster.

On the other hand, two-part codes have the advantage of being reasonably easy to understand.

Often they are also efficiently computable.
Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? We can’t encode all continuous values with finite code-lengths!
Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? We can’t encode all continuous values with finite code-lengths!

**Solution:** Quantization. Choose a discrete subset of points, \( \theta^{(1)}, \theta^{(2)}, \ldots \), and use only them.
Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? We can’t encode all continuous values with finite code-lengths!

**Solution:** Quantization. Choose a discrete subset of points, \( \theta^{(1)}, \theta^{(2)}, \ldots \), and use only them.
Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? We can’t encode all continuous values with finite code-lengths!

**Solution:** Quantization. Choose a discrete subset of points, \( \theta^{(1)}, \theta^{(2)}, \ldots \), and use only them.
Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? We can’t encode all continuous values with finite code-lengths!

**Solution:** Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.
Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? We can’t encode all continuous values with finite code-lengths!

**Solution: Quantization.** Choose a discrete subset of points, \( \theta^{(1)}, \theta^{(2)}, \ldots \), and use only them.
Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? We can’t encode all continuous values with finite code-lengths!

**Solution:** Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.
Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? We can’t encode all continuous values with finite code-lengths!

**Solution:** Quantization. Choose a discrete subset of points, $\theta^{(1)}, \theta^{(2)}, \ldots$, and use only them.

If the points are sufficiently *dense* (in a code-length sense) then the code-length for data is still almost as short as $\min_{\theta \in \Theta} \ell_\theta(D)$. 

---

*Jyrki Kivinen*  
Information-Theoretic Modeling
Continuous Parameters

What if the parameters are continuous (like polynomial coefficients)? We can’t encode all continuous values with finite code-lengths!

**Solution:** Quantization. Choose a discrete subset of points, \( \theta^{(1)}, \theta^{(2)}, \ldots \), and use only them.

If the points are sufficiently *dense* (in a code-length sense) then the code-length for data is still almost as short as \( \min_{\theta \in \Theta} \ell_\theta(D) \).
About Quantization

How many points should there be in the subset $\theta^{(1)}, \theta^{(2)}, \ldots$?
About Quantization

How many points should there be in the subset $\theta^{(1)}, \theta^{(2)}, \ldots$?

**Intuition:** Data does not allow us to tell apart $\theta_1$ and $\theta_2$ if $|\theta_1 - \theta_2| < c \frac{1}{\sqrt{n}}$. $\Rightarrow$ Don’t care about higher precision.
About Quantization

How many points should there be in the subset \( \theta^{(1)}, \theta^{(2)}, \ldots \)?

**Intuition:** Data does not allow us to tell apart \( \theta_1 \) and \( \theta_2 \) if

\[
|\theta_1 - \theta_2| < c \frac{1}{\sqrt{n}}.
\]

\( \Rightarrow \) Don’t care about higher precision.

**Theorem**

Optimal quantization accuracy is of order \( \frac{1}{\sqrt{n}} \).

\( \Rightarrow \) number of points \( \approx \sqrt{n^k} = n^{k/2} \), where \( k = \dim(\Theta) \).
How many points should there be in the subset $\theta^{(1)}, \theta^{(2)}, \ldots$?

**Intuition:** Data does not allow us to tell apart $\theta_1$ and $\theta_2$ if $|\theta_1 - \theta_2| < c \frac{1}{\sqrt{n}}$. $\Rightarrow$ Don’t care about higher precision.

**Theorem**

Optimal quantization accuracy is of order $\frac{1}{\sqrt{n}}$.

$\Rightarrow$ number of points $\approx \sqrt{n^k} = n^{k/2}$, where $k = \text{dim}(\Theta)$. 
About Quantization

How many points should there be in the subset \( \theta^{(1)}, \theta^{(2)}, \ldots \)?

**Intuition:** Data does not allow us to tell apart \( \theta_1 \) and \( \theta_2 \) if \( |\theta_1 - \theta_2| < c \frac{1}{\sqrt{n}} \). \( \Rightarrow \) Don’t care about higher precision.

**Theorem**

Optimal quantization accuracy is of order \( \frac{1}{\sqrt{n}} \).

\( \Rightarrow \) number of points \( \approx \sqrt{n^k} = n^{k/2} \), where \( k = \text{dim}(\Theta) \).

The code-length for the quantized parameters becomes

\[ \ell(\theta^q) \approx \log_2 n^{k/2} = \frac{k}{2} \log_2 n . \]
Asymptotics: $\frac{k}{2} \log n$

With the precision $\frac{1}{\sqrt{n}}$ the code-length for data is almost optimal:

$$\min_{\theta^q \in \{\theta^{(1)}, \theta^{(2)}, \ldots\}} \ell_{\theta^q}(D) \approx \min_{\theta \in \Theta} \ell_{\theta}(D) = \log_2 \frac{1}{p_\theta(D)}.$$
Asymptotics: $\frac{k}{2} \log n$

With the precision $\frac{1}{\sqrt{n}}$ the code-length for data is almost optimal:

$$\min_{\theta^q \in \{\theta^{(1)}, \theta^{(2)}, \ldots\}} \ell_{\theta^q}(D) \approx \min_{\theta \in \Theta} \ell_\theta(D) = \log_2 \frac{1}{p_\hat{\theta}(D)}. $$

The total code-length becomes then ($\approx$)

$$\log_2 \frac{1}{p_\hat{\theta}(D)} + \frac{k}{2} \log_2 n,$$

so that the regret is $\frac{k}{2} \log_2 n$. 

Jyrki Kivinen  Information-Theoretic Modeling
Asymptotics: $\frac{k}{2} \log n$

With the precision $\frac{1}{\sqrt{n}}$ the code-length for data is almost optimal:

$$\min_{\theta_q \in \{\theta^{(1)}, \theta^{(2)}, \ldots\}} \ell_{\theta_q}(D) \approx \min_{\theta \in \Theta} \ell_{\theta}(D) = \log_2 \frac{1}{p_\hat{\theta}(D)}.$$  

The total code-length becomes then ($\approx$)

$$\log_2 \left( \frac{1}{p_\hat{\theta}(D)} \right) + \frac{k}{2} \log_2 n,$$

so that the regret is $\frac{k}{2} \log_2 n$.

Since $\log_2 n$ grows slower than $n$, the **two-part code is universal** also for continuous parameter models.
1 Universal Source Codes
   • Definitions
   • Universal Models

2 Two-Part Codes
   • Discrete Parameters
   • Continuous Parameters
   • Asymptotics: $\frac{k}{2} \log n$

3 Advanced Universal Codes
   • Mixture Codes
   • Normalized Maximum Likelihood
   • Universal Prediction
Mixture Universal Model

There are universal codes that are strictly better than the two-part code.
Mixture Universal Model

There are universal codes that are strictly better than the two-part code.

For instance, given a uniquely decodable code for the parameters, let $w$ be a distribution over the parameter space $\Theta$ (quantized if necessary) defined as

$$w(\theta) = \frac{2^{-\ell(\theta)}}{c},$$

where $c = \sum_{\theta \in \Theta} 2^{-\ell(\theta)} \leq 1$. 
Mixture Universal Model

There are universal codes that are strictly better than the two-part code.

For instance, given a uniquely decodable code for the parameters, let \( w \) be a distribution over the parameter space \( \Theta \) (quantized if necessary) defined as

\[
w(\theta) = \frac{2^{-\ell(\theta)}}{c}, \quad \text{where } c = \sum_{\theta \in \Theta} 2^{-\ell(\theta)} \leq 1.
\]

Let \( p^w \) be a **mixture distribution** over the data-sets \( D \in \mathcal{D} \), defined as

\[
p^w(D) = \sum_{\theta \in \Theta} p_\theta(D) w(\theta),
\]

i.e., an “average” distribution, where each \( p_\theta \) is weighted by \( w(\theta) \).
The code-length of the mixture model $p^w$ is given by

$$\log_2 \frac{1}{\sum_{\theta \in \Theta} p_\theta(D) w(\theta)} \leq \log_2 \frac{1}{p_{\hat{\theta}}(D) w(\hat{\theta})} = \log_2 \frac{1}{p_{\hat{\theta}}(D)} + \log_2 \frac{c}{2^{-\ell(\hat{\theta})}}.$$
Mixture Universal Model

The code-length of the mixture model $p^w$ is given by

$$\log_2 \frac{1}{\sum_{\theta \in \Theta} p_\theta(D) w(\theta)} \leq \log_2 \frac{1}{p_{\hat{\theta}}(D) w(\hat{\theta})} = \log_2 \frac{1}{p_{\hat{\theta}}(D)} + \log_2 \frac{c}{2^{-\ell(\hat{\theta})}}.$$

The right-hand side is equal to

$$\log_2 \frac{1}{p_{\hat{\theta}}(D)} + \ell(\hat{\theta}) - \log_2 \frac{1}{c} \leq 0,$$

two-part code
The code-length of the mixture model \( p^w \) is given by

\[
\log_2 \left( \frac{1}{\sum_{\theta \in \Theta} p_\theta(D) w(\theta)} \right) \leq \log_2 \left( \frac{1}{p_{\hat{\theta}}(D) w(\hat{\theta})} \right) = \log_2 \frac{1}{p_{\hat{\theta}}(D)} + \log_2 \frac{c}{2^{-\ell(\hat{\theta})}}.
\]

The right-hand side is equal to

\[
\log_2 \frac{1}{p_{\hat{\theta}}(D)} + \ell(\hat{\theta}) - \log_2 \frac{1}{c},
\]

underlined \( \leq 0 \),

The mixture code is always at least as good as the two-part code.
Consider again the maximum likelihood model

\[ p_\hat{\theta}(D) = \max_{\theta \in \Theta} p_\theta(D). \]

It is the best probability assignment achievable under model \( \mathcal{M} \).
Consider again the maximum likelihood model

\[ p_{\hat{\theta}}(D) = \max_{\theta \in \Theta} p_{\theta}(D) . \]

It is the best probability assignment achievable under model \( \mathcal{M} \).

Unfortunately, it is not possible to use the ML model for coding because is not a probability distribution, i.e.,

\[ C = \sum_{D \in \mathcal{D}} p_{\hat{\theta}}(D) > 1 , \]

unless \( \hat{\theta} \) is constant wrt. \( D \).
The normalized maximum likelihood (NML) model is obtained by normalizing the ML model:

$$p_{nml}(D) = \frac{p_{\hat{\theta}}(D)}{C}, \quad \text{where } C = \sum_{D \in \mathcal{D}} p_{\hat{\theta}}(D).$$
Normalized Maximum Likelihood

The **normalized maximum likelihood (NML) model** is obtained by normalizing the ML model:

\[
p_{\text{nml}}(D) = \frac{p_{\hat{\theta}}(D)}{C}, \quad \text{where } C = \sum_{D \in \mathcal{D}} p_{\hat{\theta}}(D).
\]

The regret of NML is given by

\[
\log_2 \frac{1}{p_{\text{nml}}(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)} = \log_2 \frac{C}{p_{\hat{\theta}}(D)} - \log_2 \frac{1}{p_{\hat{\theta}}(D)} = \log_2 C,
\]

which is constant wrt. \(D\).
Model Complexity

The quantity $\log_2 C$, which gives the (constant) regret of NML, is called the *complexity* of model class $\mathcal{M}$. 
Model Complexity

The quantity $\log_2 C$, which gives the (constant) regret of NML, is called the *complexity* of model class $\mathcal{M}$.

Notice that if $\mathcal{D}$ and $\mathcal{M}$ are infinite, the sum defining $C$ may diverge. In this case, we say that the complexity of the model is infinite.
Model Complexity

The quantity $\log_2 C$, which gives the (constant) regret of NML, is called the *complexity* of model class $\mathcal{M}$.

Notice that if $\mathcal{D}$ and $\mathcal{M}$ are infinite, the sum defining $C$ may diverge. In this case, we say that the complexity of the model is infinite.

If the complexity is infinite, then it’s impossible to achieve constant regret. This is a real issue for many (but not all) model classes used in practice.
The quantity $\log_2 C$, which gives the (constant) regret of NML, is called the \textit{complexity} of model class $\mathcal{M}$.

Notice that if $\mathcal{D}$ and $\mathcal{M}$ are infinite, the sum defining $C$ may diverge. In this case, we say that the complexity of the model is infinite.

If the complexity is infinite, then it’s impossible to achieve constant regret. This is a real issue for many (but not all) model classes used in practice.

Various work-arounds exist to extend NML to model classes with infinite complexity.
Let $q$ be any distribution other than $p_{\text{nml}}$. Then

- there must a data-set $D' \in \mathcal{D}$ for which we have

$$q(D') < p_{\text{nml}}(D')$$
Normalized Maximum Likelihood

Let $q$ be any distribution other than $p_{nml}$. Then

- there must be a data-set $D' \in \mathcal{D}$ for which we have

\[
q(D') < p_{nml}(D')
\]

\[
\Leftrightarrow \log_2 \frac{1}{q(D')} - \log_2 \frac{1}{p_{\hat{\theta}}(D')} > \log_2 \frac{1}{p_{nml}(D')} - \log_2 \frac{1}{p_{\hat{\theta}}(D')}
\]

\[
\text{regret of } q \quad \text{regret of } p_{nml}
\]
Let $q$ be any distribution other than $p_{\text{nml}}$. Then

- there must a data-set $D' \in \mathcal{D}$ for which we have

\[
q(D') < p_{\text{nml}}(D')
\]

\[
\Leftrightarrow \log_2 \frac{1}{q(D')} - \log_2 \frac{1}{p_{\hat{\theta}}(D')} > \log_2 \frac{1}{p_{\text{nml}}(D')} - \log_2 \frac{1}{p_{\hat{\theta}}(D')} ,
\]

regret of $q$

regret of $p_{\text{nml}}$

For $D'$, the regret of $q$ is greater than $\log_2 C$, the regret of $p_{\text{nml}}$. 
Let $q$ be any distribution other than $p_{\text{nml}}$. Then
- there must be a data-set $D' \in \mathcal{D}$ for which we have

$$q(D') < p_{\text{nml}}(D')$$

$$\iff \log_2 \frac{1}{q(D')} - \log_2 \frac{1}{p_{\hat{\theta}}(D')} > \log_2 \frac{1}{p_{\text{nml}}(D')} - \log_2 \frac{1}{p_{\hat{\theta}}(D')} ,$$

regret of $q$

regret of $p_{\text{nml}}$

For $D'$, the regret of $q$ is greater than $\log_2 C$, the regret of $p_{\text{nml}}$.

Thus, the worst-case regret of $q$ is greater than the (worst-case) regret of NML. $\Rightarrow$ NML has the least possible worst-case regret.
Universal Models

For ‘smooth’ parametric models, the regret of NML, $\log_2 C$, grows slower than $n$, so **NML is also a universal model.**
Universal Models

For ‘smooth’ parametric models, the regret of NML, $\log_2 C$, grows slower than $n$, so NML is also a universal model.

We have seen three kinds of universal codes:

1. two-part,
2. mixture,
3. NML.
For ‘smooth’ parametric models, the regret of NML, $\log_2 C$, grows slower than $n$, so **NML is also a universal model.**

We have seen three kinds of universal codes:

1. two-part,
2. mixture,
3. NML.

There are also universal codes that are not based on any (explicit) model class: Lempel-Ziv (gzip)!
So what do we do with them?
So what do we do with them?

We can use universal codes for (at least) three purposes:
So what do we do with them?

We can use universal codes for (at least) three purposes:

1. compression,
So what do we do with them?

We can use universal codes for (at least) three purposes:

1. compression,
2. prediction,
So what do we do with them?

We can use universal codes for (at least) three purposes:

1. compression,
2. prediction,
3. model selection.
Universal Prediction

By the connection \( p(D) = 2^{-\ell(D)} \), the following are equivalent:

- **good compression:** \( \ell(D) \) is small,
Universal Prediction

By the connection \( p(D) = 2^{-\ell(D)} \), the following are equivalent:

- **good compression**: \( \ell(D) \) is small,
- **good probability assignment**: 
  \[
  p(D) = \prod_{i=1}^{n} P(D_i | D_1, \ldots, D_{i-1}) \text{ is high.}
  \]
By the connection $p(D) = 2^{-\ell(D)}$, the following are equivalent:

- **good compression**: $\ell(D)$ is small,
- **good probability assignment**: 
  $$p(D) = \prod_{i=1}^{n} P(D_i \mid D_1, \ldots, D_{i-1})$$ is high.
- **good predictions**: 
  $$p(D_i \mid D_1, \ldots, D_{i-1})$$ is high for most 
  $i \in \{1, \ldots, n\}$. 

For instance, the mixture code gives a natural predictor which is equivalent to Bayesian prediction. The NML model gives predictions that are good relative to the best model in the model class, no matter what happens.
Universal Prediction

By the connection $p(D) = 2^{-\ell(D)}$, the following are equivalent:

- **good compression**: $\ell(D)$ is small,
- **good probability assignment**: $p(D) = \prod_{i=1}^{n} P(D_i \mid D_1, \ldots, D_{i-1})$ is high.
- **good predictions**: $p(D_i \mid D_1, \ldots, D_{i-1})$ is high for most $i \in \{1, \ldots, n\}$.

For instance, the mixture code gives a natural predictor which is equivalent to **Bayesian prediction**.
Universal Prediction

By the connection $p(D) = 2^{-\ell(D)}$, the following are equivalent:

- **good compression:** $\ell(D)$ is small,
- **good probability assignment:**
  
  $$p(D) = \prod_{i=1}^{n} P(D_i \mid D_1, \ldots, D_{i-1})$$
  is high.
- **good predictions:** $p(D_i \mid D_1, \ldots, D_{i-1})$ is high for most $i \in \{1, \ldots, n\}$.

For instance, the mixture code gives a natural predictor which is equivalent to **Bayesian prediction**.

The NML model gives predictions that are good relative to the best model in the model class, **no matter what happens**.
Since a model class that enables good compression of the data must be based on exploiting the regular features in the data, the code-length can be used as a yard-stick for comparing model classes.
**MDL Principle**

**“Old-style”:**
- Choose the model $p_\theta \in \mathcal{M}$ that yields the shortest *two-part code-length*

$$\min_{\theta, \mathcal{M}} \ell(\mathcal{M}) + \ell_1(\theta) + \log_2 \frac{1}{p_\theta(D)}.$$

**Modern:**
- Choose the model class $\mathcal{M}$ that yields the shortest *universal code-length*

$$\min_{\mathcal{M}} \ell(\mathcal{M}) + \ell_\mathcal{M}(D).$$
Next week: Minimum Description Length (MDL) principle