Lecture 2, Part 1:

# Theoretical Bounds for Lossless Compression

Robert Bamler · Summer Term of 2023

These slides are part of the course *"Data Compression With and Without Deep Probabilistic Models"* taught at University of Tübingen. More course materials—including video recordings, lecture notes, and problem sets with solutions—are publicly available at `https://robamler.github.io/teaching/compress23/`.

## Admin Stuff



- **Important:** next lecture *only on zoom*, not in classroom
  - Sign up to course using (new) Ilias link to get zoom link by email
    (link will also be on website ∼30 minutes before next week's lecture starts)
- You'll have to sign up for *exam* on Alma starting 5 June (independently of whether you signed up to the *course* on Ilias)
  - More details will follow.



## Recap: Symbol Codes



i.e., finite or countably infinite

- *alphabet* $\mathfrak{X}$ (discrete set) with *probabilities* $p(x)$ for all *symbols* $x \in \mathfrak{X}$
- *message* $\mathbf{x} = (x_1, x_2, \ldots, x_{k(\mathbf{x})}) \in \mathfrak{X}^*$
- *code book* $C$ maps any $x \in \mathfrak{X}$ to its *code word* $C(x) \in \{0, \ldots, B-1\}^*$ (usually: $B = 2$)
  - induces a *symbol code* $C^*\colon \mathfrak{X}^* \to \{0, \ldots, B-1\}^*$ by concatenation (without delimiters):
    $C^*(\mathbf{x}) := C(x_1) \,\|\, C(x_2) \,\|\, \ldots \,\|\, C(x_{k(\mathbf{x})})$
- properties of symbol codes:
  - *unique decodability:* $C^*$ is injective
  - *prefix code:* no code word $C(x)$ is a prefix of another code word $C(x')$ with $x' \neq x$
  - $C$ is a prefix code $\Rightarrow$ $C$ is uniquely decodable (but reverse is in general not true) → Problem 0.2 (d)
- *expected code word length* $L_C := \sum_{x \in \mathfrak{X}} p(x)\,|C(x)|$
  ← (simple) model of the data source (crucial for source coding)
  → Problem set 1
- *Huffman coding* generates an optimal symbol code (that minimizes $L_C$) for a given $p$

## Theoretical Bounds for Lossless Compression

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- **Goal of this lecture:** Source Coding Theorem [Shannon, 1948]
  - Relates $L_C$ to the so-called *entropy* $H_B[p]$ (which we'll define later today).
  - **The Bad News:** a uniquely decodable $B$-ary symbol code $C$ *cannot* have $L_C < H_B[p]$.
  - **The Good News:** $\forall p$, one can make $L_C$ close to $H_B[p]$ with less than 1 bit per symbol overhead.
- **Step 1:** proof bound on code word lengths, *independent of* $p$ (KM-Theorem)
- **Step 2:** proof bound on *expected* code word length for a given model $p$
- **Credits:** Our proof follows:
  `https://www.youtube.com/watch?v=yHw1ka-4g0s&list=PLE125425EC837021F&index=14`



## The Kraft-McMillan Theorem [Kraft, 1949; McMillan, 1956]

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

(a) $\forall$ $B$-ary uniquely decodable symbol codes over some discrete alphabet $\mathfrak{X}$:

$$\sum_{x\in\mathfrak{X}} \frac{1}{B^{|C(x)|}} \leq 1 \qquad \text{("Kraft inequality").} \tag{1}$$

**Interpretation:** we have a finite budget of "shortness" for code words:

- interpret $\frac{1}{B^{|C(x)|}}$ as the "shortness" of code word $C(x)$;
- the sum of all "shortnesses" must not exceed 1;
- if we shorten one code word then we may have to make another code word longer so that we don't exceed our "shortness budget".

Assume you are given a uniquely decodable symbol code and want to shorten the code word for some specific symbol $x \in \mathfrak{X}$.
$\Rightarrow |C(x)|$ shrinks
$\Rightarrow$ "shortness" $B^{-|C(x)|}$ grows
$\Rightarrow$ to avoid exceeding your "shortness budget", you must reduce the shortness $B^{-|C(x')|}$ of some $x' \neq x$
$\Rightarrow |C(x')|$ grows

(b) $\forall$ functions $\ell : \mathfrak{X} \to \mathbb{N}$ that satisfy the Kraft inequality (i.e., $\sum_{x\in\mathfrak{X}} \frac{1}{B^{\ell(x)}} \leq 1$):

$$\exists \text{ } B\text{-ary prefix code } C_\ell \text{ with } |C_\ell(x)| = \ell(x) \quad \forall x \in \mathfrak{X}.$$

(combining parts (a) and (b))

**Corollary:** $\forall$ uniquely decodable $B$-ary symbol codes $C$:
$\exists$ a $B$-ary *prefix* code $C'$ with same code word lengths (i.e., $|C'(x)| = |C(x)|$ $\forall x \in \mathfrak{X}$)



## Lemma

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- let: $\begin{cases} C \text{ be a } B\text{-ary uniquely decodable symbol code over } \mathfrak{X}; \\ s \in \mathbb{N}_0; \\ Y_s := \{\mathbf{x} \in \mathfrak{X}^* \text{ with } |C^*(\mathbf{x})| = s\}. \end{cases}$
- then: $|Y_s| \leq B^s$.

**Proof:** Let $S_s := \{ \underbrace{C^*(\underline{x})}_{\in \{0,\dots,B-1\}^s} : \underline{x} \in Y_s \} \subseteq \underbrace{\{0,\dots,B-1\}^s}_{\text{size: } B^s} \Rightarrow |S_s| \leq B^s$

Assume $|Y_s| > B^s \Rightarrow |Y_s| > |S_s|$

$\Rightarrow \exists\, \underline{x}, \underline{x}' \in Y_s$ with $\underline{x} \neq \underline{x}'$ but $C^*(\underline{x}) = C^*(\underline{x}')$

$\Rightarrow C^*$ not injective, i.e., $C$ not uniquely decodable

# Proof of Part (a) of KM Theorem



**Claim (reminder):** $C$ is uniquely decodable $\implies \underbrace{\sum_{x\in\mathfrak{X}} \frac{1}{B^{|C(x)|}}}_{=:r} \le 1.$

Proof: Let $k \in \mathbb{N}$.

$$r^k = \Big(\sum_{x\in\mathfrak{X}} B^{-|C(x)|}\Big)^k = \Big(\sum_{x_1\in\mathfrak{X}} B^{-|C(x_1)|}\Big)\Big(\sum_{x_2\in\mathfrak{X}} B^{-|C(x_2)|}\Big)\cdots\Big(\sum_{x_k\in\mathfrak{X}} B^{-|C(x_k)|}\Big) = \sum_{\underline{x}\in\mathfrak{X}^k} B^{-|C^*(\underline{x})|}$$

(i) if $\mathfrak{X}$ is finite: $\Rightarrow \gamma := \max_{x\in\mathfrak{X}} |C(x)| < \infty$ is well-defined & finite;

$$r^k = \sum_{\underline{x}\in\mathfrak{X}^k} B^{-|C^*(\underline{x})|} = \sum_{s=0}^{\gamma k}\sum_{\underline{x}\in Y_s} B^{-s} = \sum_{s=0}^{\gamma k} \underbrace{|Y_s|}_{\le B^{s}\ \text{(Lemma)}} B^{-s} \le \gamma k + 1 \Rightarrow \underbrace{\frac{r^k-1}{k}}_{\to\infty \text{ for } k\to\infty \text{ if } r>1} < \underbrace{\gamma}_{\text{indep. of } k}$$

$\Rightarrow r \le 1$

(ii) if $\mathfrak{X}$ is countably infinite: without restriction, assume $\mathfrak{X} = \mathbb{N}$;

$$\Rightarrow r = \sum_{x\in\mathbb{N}} B^{-|C(x)|} = \sum_{x=1}^{\infty} B^{-|C(x)|} = \lim_{n\to\infty} \underbrace{\sum_{x=1}^{n} B^{-|C(x)|}}_{\text{alphabet of size } n<\infty \Rightarrow \text{case (i) applies}} \le 1$$

all terms $\ge 0$ $\Rightarrow$ absolutely convergent if convergent at all $\Rightarrow$ may reorder terms arbitrarily



# Proof of Part (b) of KM Theorem



output of algorithm below

**Claim (reminder):** $\sum_{x\in\mathfrak{X}} \frac{1}{B^{\ell(x)}} \le 1 \implies \exists$ $B$-ary prefix code $C_\ell$ with $|C_\ell(x)| = \ell(x)\ \forall x \in \mathfrak{X}$.

input of algorithm below

**Constructive proof:** we show existence of $C$ by showing how it can be obtained.

Algorithm: sort symbols in $\mathfrak{X} = \{x, x', x'', \ldots\}$ s.t. $\ell(x) \ge \ell(x') \ge \ell(x'') \ge \ldots$;
initialize $\xi \leftarrow 1$;
for each $x \in \mathfrak{X}$ in above order:
- update $\xi \leftarrow \xi - B^{-\ell(x)}$;
- write $\xi \in [0,1)$ in $B$-ary: $\xi = (0.\underbrace{????}\ldots)_B$;
- set $C(x)$ to first $\ell(x)$ bits here (pad with trailing zeros if necessary)

also discusses case of (countably) infinite $\mathfrak{X}$

**Claim:** The resulting code book $C_\ell$ is prefix free (proof: Problem 2.1).



# Example: Simplified Game of Monopoly (SGoM)



| $x$ | $\ell(x)$ | $C_\ell(x)$ | order after sorting by descending $\ell(x)$ |
|---|---|---|---|
| 2 | 3 | 111 | 1 |
| 3 | 2 | 10 | 3 |
| 4 | 2 | 01 | 4 |
| 5 | 2 | 00 | 5 |
| 6 | 3 | 110 | 2 |

$\xi \leftarrow 1$ (initialization)

$\xi \leftarrow 1 - 2^{-3} = (1.000)_2 - (0.001)_2 = (0.111)_2$

$\xi \leftarrow (0.110)_2 - (0.01)_2 = (0.10)_2$

$\xi \leftarrow (0.10)_2 - (0.01)_2 = (0.01)_2$

$\xi \leftarrow (0.01)_2 - (0.01)_2 = (0.00)_2$

$\xi \leftarrow (0.111)_2 - (0.001)_2 = (0.110)_2$

exercise: execute algorithm without sorting $\mathfrak{X}$ by descending $\ell(x)$ and verify that it fails.

- Check Kraft inequality for $B = 2$: $\sum_{x\in\mathfrak{X}} 2^{-\ell(x)} = 2\times 3^{-2} + 3\times 2^{-2} = 1 \le 1$ ✓
- **Question:** how should we choose $\ell : \mathfrak{X} \to \mathbb{N}$ for a given probabilistic model $p$?
  - **optimally:** via Huffman coding
  - **near-optimally:** via *information content* (next part).

# Outlook

- **Problem Set 2:**
  - complete proof of part (b) of KM-Theorem
  - implement Huffman *decoder* in Python
- **Next part:**
  - theoretical bounds on the expected code word length $L_C$ ("The Bad News" & "The Good News")
  - theoretical bounds *beyond symbol codes:* Source Coding Theorem



EBERHARD KARLS UNIVERSITÄT TÜBINGEN
Faculty of Science · Department of Computer Science · Group of Prof. Robert Bamler

Lecture 2, Part 2:

# The Source Coding Theorem

Robert Bamler · Summer Term of 2023

These slides are part of the course *"Data Compression With and Without Deep Probabilistic Models"* taught at University of Tübingen. More course materials—including video recordings, lecture notes, and problem sets with solutions—are publicly available at `https://robamler.github.io/teaching/compress23/`.

# Recap: Kraft-McMillan (KM) Theorem

(a) $\forall$ $B$-ary uniquely decodable symbol codes over some discrete alphabet $\mathfrak{X}$:

$$\sum_{x\in\mathfrak{X}} \frac{1}{B^{|C(x)|}} \leq 1 \qquad (\text{"Kraft inequality"}). \tag{1}$$

(b) $\forall$ functions $\ell : \mathfrak{X} \to \mathbb{N}$ that satisfy the Kraft inequality (i.e., $\sum_{x\in\mathfrak{X}} \frac{1}{B^{\ell(x)}} \leq 1$):

$$\exists\ B\text{-ary prefix code } C_\ell \text{ with } |C_\ell(x)| = \ell(x)\ \forall x \in \mathfrak{X}.$$

- **Question:** how should we choose $\ell : \mathfrak{X} \to \mathbb{N}$ for a given probabilistic model $p$?
  - **optimally:** via Huffman coding (problem: no closed-form solution)
  - **near-optimally (this part):** via *information content*
    **spoiler:** $\ell_S(x) := \lceil -\log_B p(x) \rceil$

## Optimal Choice of $\ell$

- **Constrained optimization problem:** $(\star)$
  - minimize: $L_{C_\ell} = \sum_{x\in\mathfrak{X}} p(x)\,|C_\ell(x)| = \sum_{x\in\mathfrak{X}} p(x)\,\ell(x)$
  - constraints: (i) $\sum_{x\in\mathfrak{X}} \frac{1}{B^{\ell(x)}} \leq 1$; (ii) $\ell(x) \in \mathbb{N}\ \ \forall x \in \mathfrak{X}$.
- **Idea:** relax constraint (ii): $(\square)$
  - minimize: $L_\ell := \sum_{x\in\mathfrak{X}} p(x)\,\ell(x)$
  - constraints: (i) $\sum_{x\in\mathfrak{X}} \frac{1}{B^{\ell(x)}} \leq 1$; (ii') $\ell(x) \in \mathbb{R}_{>0}\ \ \forall x \in \mathfrak{X}$.

  $\Rightarrow$ yields *lower bound*: solution $L_\ell$ of $(\square) \leq$ solution $L_{C_\ell}$ of $(\star)$
- **Observation:** solution of $(\square)$ satisfies: (i') $\sum_{x\in\mathfrak{X}} \frac{1}{B^{\ell(x)}} = 1$.
  - Enforce via Lagrange multiplier $\lambda$:
    find stationary point of $\mathcal{L}_{\ell,\lambda} := L_\ell + \lambda\big(\sum_{x\in\mathfrak{X}} \frac{1}{B^{\ell(x)}} - 1\big)$ w.r.t. $\lambda \in \mathbb{R}$ and all $\ell(x) \in \mathbb{R}_{\geq 0}\ \forall x \in \mathfrak{X}$.

- $0 = \frac{\partial \mathcal{L}_{\ell,\lambda}}{\partial \lambda} = \sum_{x\in\mathfrak{X}} B^{-\ell(x)} - 1 \Leftrightarrow$ constraint (ii')
- $\forall x \in \mathfrak{X}$: $0 = \frac{\partial \mathcal{L}_{\ell,\lambda}}{\partial \ell(x)} = \frac{\partial L_\ell}{\partial \ell(x)} + \lambda \frac{\partial}{\partial \ell(x)} B^{-\ell(x)}$

$$= p(x) + \lambda \frac{\partial}{\partial \ell(x)} e^{\ln(B^{-\ell(x)})}$$
$$= p(x) + \lambda \frac{\partial}{\partial \ell(x)} e^{-\ell(x)\ln B}$$
$$= p(x) - \lambda \ln B\, e^{-\ell(x)\ln B}$$
$$= p(x) - \lambda \ln B\, B^{-\ell(x)}$$

- solve for $\ell(x)$:

$$\ell(x) = -\log_B\left(\frac{p(x)}{\lambda \ln B}\right)$$
$$= -\log_B p(x) + \alpha \quad \alpha = \log_B(\lambda \ln B)$$

- obtain $\lambda$ from constraint:

$$1 = \sum_{x\in\mathfrak{X}} B^{-\ell(x)} = B^{-\alpha} \underbrace{\sum_{x\in\mathfrak{X}} p(x)}_{=1} \Rightarrow \underline{\alpha = 0}$$





## Lower Bound on Expected Code Word Length $L_C$

- **Solution of relaxed optimization problem $(\square)$:** $\ell(x) = \underbrace{-\log_B p(x)}_{\text{"information content of the symbol } x\text{"}}$ (under model $p$ and to base $B$)
  - $L_\ell = \sum_{x\in\mathfrak{X}} p(x)\,\ell(x) = \underbrace{-\sum_{x\in\mathfrak{X}} p(x) \log_B p(x)}_{=: H_B[p] \text{ ("entropy")}}$
- Let's now restore the constraints from $(\star)$, i.e., $\ell : \mathfrak{X} \to \mathbb{N}$ must be *integer valued*.
  - **Recall:** solution $L_{C_\ell}$ of $(\star) \geq$ solution $L_\ell$ of $(\square)$
  - Thus, for all *integer valued* $\ell$ that satisfy Kraft inequality: $L_{C_\ell} \geq H_B[p]$
- By part (a) of the KM-Theorem:

**lower bound on the expected code word length $L_C$ of any uniquely decodable $B$-ary symbol code $C$:**
$$L_C \geq H_B[p]$$





## Shannon Coding [Shannon, 1948]

- **Last slide:** "the bad news"
  - Lower bound for uniquely decodable $B$-ary symbol code: $L_C \geq H_B[p] = -\sum_{x\in\mathfrak{X}} p(x) \log_B p(x)$
  - We would achieve equality ($L_C = H_B[p]$) if we were able to set $\ell(x) = \underbrace{-\log_B p(x)}_{\notin \mathbb{N} \text{ (in general)}}\ \ \forall x \in \mathfrak{X}$.
- **Question:** How closely can we approach this bound?
- **Idea:** choose $\ell_S : \mathfrak{X} \to \mathbb{N}$ as follows: $\ell_S(x) = \lceil -\log_B p(x) \rceil$ — $\lceil\cdots\rceil$ denotes rounding up to nearest integer
  - Satisfies Kraft inequality: $\sum_{x\in\mathfrak{X}} B^{-\ell_S(x)} = \sum_{x\in\mathfrak{X}} B^{-\lceil -\log_B p(x) \rceil} \leq \sum_{x\in\mathfrak{X}} B^{\log_B p(x)} = \sum_{x\in\mathfrak{X}} p(x) = 1$
- **By part (b) of KM-Theorem:** $\exists$ $B$-ary prefix code $C_S$ with $|C_S(x)| = \ell_S(x)\ \ \forall x \in \mathfrak{X}$.
  - $L_{C_S} = \sum_{x\in\mathfrak{X}} p(x)\,\ell_S(x) = \sum_{x\in\mathfrak{X}} p(x) \lceil -\log_B p(x) \rceil < \sum_{x\in\mathfrak{X}} p(x)\big(-\log_B p(x) + 1\big) = H_B[p] + 1$
  - in short: $L_{C_S} < H_B[p] + 1$ "the good news"

# Symmary: Theoretical Bounds for symbol codes



- **The Bad News:** no (uniquely decodable $B$-ary) symbol code can have an expected code word length smaller than the entropy $H_B[p]$ of a symbol.
- **The Good News:** one can always approach this lower bound with less than 1 bit of overhead *per symbol* (e.g., by using the *Shannon code* $C_S$).
- Thus, the *optimal* code $C_{\text{opt}}$ (that minimizes $L_C$) satisfies:

$$\boxed{H_B[p] \le L_{C_{\text{opt}}} < H_B[p] + 1}$$

(but this requires that $|C(x')| > -\log_B p(x')$ for some $x' \neq x$, see discussion of KM-theorem)

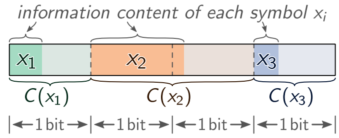
- **Note:** The above bounds are *in expectation over all symbols $x \in \mathfrak{X}$*.
  - For any *specific* symbol $x \in \mathfrak{X}$, a code $C$ can "violate the lower bound": $|C(x)| < -\log_B p(x)$.
  - But: *Shannon code* satisfies $-\log_B p(x) \le |C_S(x)| < -\log_B p(x) + 1$ *for each individual $x \in \mathfrak{X}$*.



# The Source Coding Theorem [Shannon, 1948]



- **So far:** theoretical bounds for *symbol codes*: $\boxed{H_B[p] \le L_{C_{\text{opt}}} < H_B[p] + 1}$
- **Symbol codes are suboptimal.**
  - Always generate an *integer* number of bits per symbol.
  - Thus, overhead of up to 1 bit applies *per symbol*.




- **Practical solution:** stream codes (Lectures 5 and 6)
- **For theoretical analysis:** consider entire message $\mathbf{x} \in \mathfrak{X}^*$ as a single symbol.
  - New alphabet $\mathfrak{X}^*$ is still *countable*, thus theorems still apply.
  - Probability distribution $p^*$ on $\mathfrak{X}^*$ can be complicated, but we'll assume it has a finite entropy $H_B[p^*] = -\sum_{\mathbf{x} \in \mathfrak{X}^*} p^*(\mathbf{x}) \log_B p^*(\mathbf{x})$.

infinitely large; and even if we set a maximum message length $k$ then $|\mathfrak{X}^k| = |\mathfrak{X}|^k$ is still exponentially large $\Rightarrow$ Huffman coding directly on this space would be astronomically expensive.

$\Rightarrow$ The optimal uniq. dec. code $C_{\text{opt}}$ on $\mathfrak{X}^*$ (typically *not* a symbol code on $\mathfrak{X}$) satisfies: ($\Rightarrow$ need stream codes)

$$\boxed{H_B[p^*] \le \text{expected bit rate of } C_{\text{opt}} < H_B[p^*] + 1}$$



# Outlook



- **Problem Set 2:**
  - simple examples of Shannon coding
  - entropy and information content
- **Next week (on zoom!):**
  - proof of optimality of Huffman coding
  - machine-learning models for lossless compression (continued in Lectures 4 and 7-9)
- **Lectures 5 & 6:** beyond symbol codes: stream codes
- **Lecture 11:** theoretical bounds for *lossy* compression ("Rate/Distortion Theory")