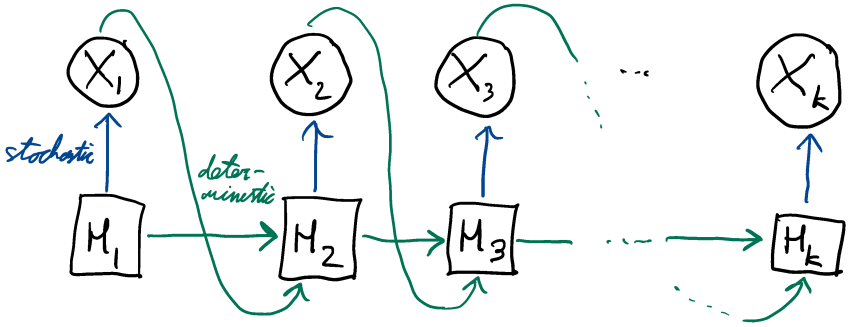
Lecture notes from the course “Data Compression With Deep Probabilistic Models” by Prof. Robert Bamler, University of Tuebingen, 24 May 2021; more course materials at https://robamler.github.io/teaching/compress21/

Bits-Back Coding With Latent Variable Models

Last video:
- overview of probability theory & random vars.
- modeling error: KL-Divergence
  → need to model correlations! → Problem 4.2 & 6.2 $H_p(X) + H_p(Y) \geq H_p((X,Y))$
  → autoregressive models



- ☺ can model long(-ish) range correlations
- ☹ struggle with very long correlations
- ☺ compact to store
- ☹ not parallelizable

This video:
- latent variable models
- Bayesian inference
- Bits-back coding

Latent Variable Models

Consider these hypothetical news headlines.

Parliament Votes on New Labor Bill.

Labor Union Votes to Extend Strikes.

} topic "politics"

Soccer Player Scores First Goal Since Joining New Team.

Guest Team is Leading by One Goal.

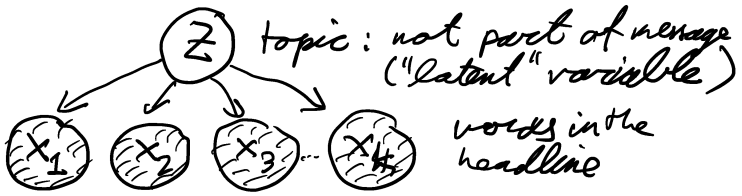
} topic "sports"

Observation: words within a headline appear to be correlated:

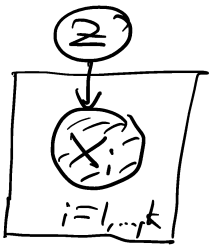
consider two positions $i \neq j$; claim: X_i & X_j are not stat. indep
(words at those positions)

e.g.: $P(X_i = \text{"Goal"}, X_j = \text{"Team"}) > P(X_i = \text{"Goal"})\, P(X_j = \text{"Team"})$

Model of the "Generative Process"



Abbreviated depiction



These pictures denote a joint prob. dist. that factorizes as follows:

$$P(\underline{X}, Z) = P(Z)\, P(\underline{X} \mid Z) = P(Z) \prod_{i=1}^{k} P(X_i \mid Z)$$

("Topic Model", e.g. LDA: Blei & Ng 2003, before in genetics Pritchard et al. 2000)

⇒ marginal dist of the message $\underline{X}$:

$$P(\underline{X}) = \sum_z P(\underline{X}, Z = z) = \sum_z \left(P(Z = z) \prod_{i=1}^{k} P(X_i \mid Z = z) \right)$$

Claim: this model can capture correlations like

$$P(X_i = \text{"Goal"}, X_j = \text{"Team"}) > P(X_i = \text{"Goal"})\, P(X_j = \text{"Team"})$$

⇒ proof: exercise

Data Compression With Latent Variable Models

$$P(\underline{X}, z) = P(z)\, P(\underline{X} \mid z)$$

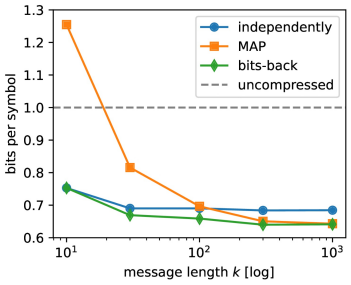
$\underline{X}$: message, z: latent var

ideally we would like to compress $\underline{X}$ with this model
→ Problem: we don't know value of z

Problem Set 5: implement & compare 3 compression methods for l.v.m.

→ **Problem 5.2:** treat X_i as independent
⇒ ignore correlations
→ $\mathbb{E}_p[R(\underline{x})] = \sum_{i=1}^{} H_p(X_i) \geq H_p(\underline{X})$

→ **Problem 5.3:** MAP-estimate
→ $\mathbb{E}_p[R(\underline{x})] = -\log P(z=z^*) + H_p(\underline{X} \mid z=z^*)$



→ **Problem 5.4:** bits-back coding
→ $\mathbb{E}_p[R_{net}(\underline{x})] = H_p(\underline{X})$

Naive approach: MAP-estimate

$$P(\underline{X}, z) = P(z)\, P(\underline{X} \mid z)$$

→ Idea:
- encode some value z for z using $P(z)$ & transmit it
- then encode $\underline{X}$ using $P(\underline{X} \mid z=z)$ $\left(= \prod_{i=1}^{k} P(X_i \mid z=z)\right)$

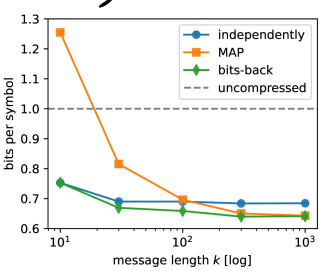
⇒ decoder can
- decode z using $P(z)$
- decode $\underline{x}$ using $P(\underline{X} \mid z=z)$

bit rate: $R^{(z)}(\underline{x}) = -\log P(z=z) - \log P(\underline{X}=\underline{x} \mid z=z)$

$$= -\log P(\underline{X}=\underline{x},\ z=z)$$

chose $z^* := \arg\min_z R^{(z)}(\underline{x}) = \arg\max_z P(\underline{X}=\underline{x},\ z=z)$

↑ "maximum a-posteriori" (MAP) estimate of z



overhead over theoret. bound:

$$R^{(z^*)}(\underline{x}) - (-\log P(\underline{X}=\underline{x})) = -\log P(\underline{X}=\underline{x},\ z=z) + \log P(\underline{X}=\underline{x})$$

$$= -\log P(z=z \mid \underline{X}=\underline{x}) \geq 0$$

← posterior distribution

Bayesian Inference

- model: $P(\underline{X}, Z) = P(Z)\, P(\underline{X} \mid Z)$

↳ know $\underline{X}$, don't know Z (⇒ MAP-estimate method has an overhead)

↳ But: knowing $\underline{X}$ typically reveals some information about Z

Parliament Votes on New Labor Bill. } topic "politics"

Labor Union Votes to Extend Strikes. }

Soccer Player Scores First Goal Since Joining New Team. } topic "sports"

Guest Team is Leading by One Goal. }

⇒ However: there can still be some ambiguity about Z (even after you know $\underline{X}$)

Parliament Votes on Aid for Community Sports Teams. (*)

⇒ can only make prob. statements about Z

⇒ posterior $P(Z \mid \underline{X} = \underline{x}) = \dfrac{P(Z)\, P(\underline{X} = \underline{x} \mid Z)}{P(\underline{X} = \underline{x})} = \dfrac{P(Z)\, P(\underline{X} = \underline{x} \mid Z)}{\sum_{z'} P(Z = z')\, P(\underline{X} = \underline{x} \mid Z = z')}$

(prior, likelihood, posterior)

Remarks:

- in principle posterior distrib. is known once you know $P(\underline{X}, Z)$ & $\underline{x}$
- in practice, however, calculating the posterior is often prohibitively expensive

(→ Lecture 7: approximate Bayesian inference)

Understanding the overhead of MAP-est. method

→ we could encode (*) in two different ways

ⓐ $Z =$ "politics", then we use $P(\underline{X} \mid Z =$ "politics"$)$

ⓑ $Z =$ "sports", then we use $P(\underline{X} \mid Z =$ "sports"$)$

two different compressed bit strings of same message → wasteful

→ remember: overhead $= -\log P(Z = z^* \mid \underline{X} = \underline{x})$

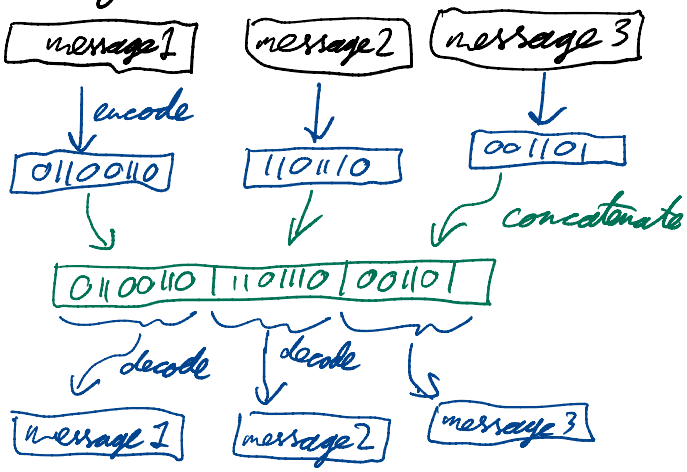
Bits-Back Coding

(Wallace 1990, Hinton & Camp 1993; practical: BB-ANS: Townsend et al. 2019 lossy: Yang et al. 2020)

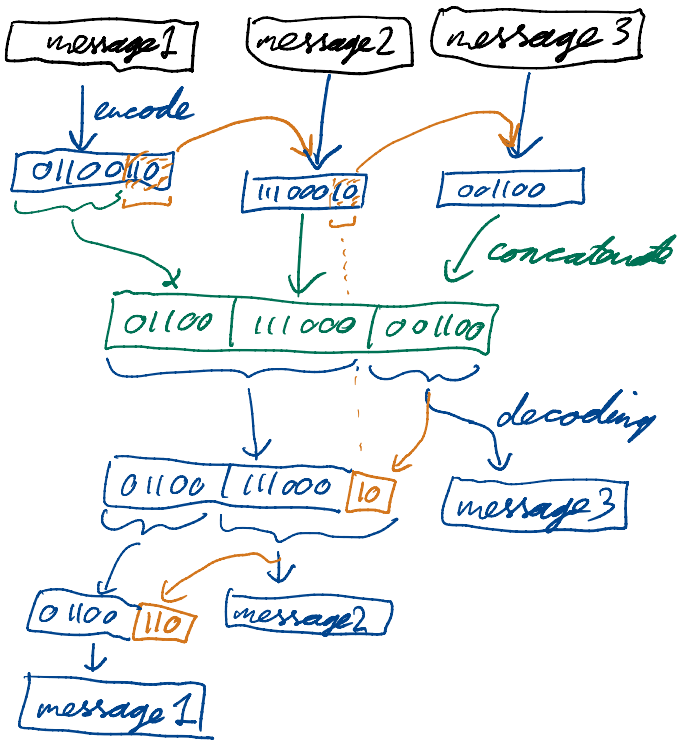
→ Idea: "piggyback" some additional message into the choice of z

Setup: communicate multiple messages (e.g. multiple image patches) over a single channel

→ usually:



→ bits-back: (operates as a stack, i.e. "last in first out")



Algorithm: "Bits-Back Coding"

subroutine encode ($\underline{x}$, compressed, P): (compressed: bit string)

- $z \leftarrow$ decode from compressed using $P(\underline{Z} \mid \underline{X}=\underline{x})$;
- encode $\underline{x}$ using $P(\underline{X} \mid \underline{Z}=z)$ onto compressed
- encode z using $P(\underline{Z})$ onto compressed
- return compressed

subroutine decode (compressed, P):

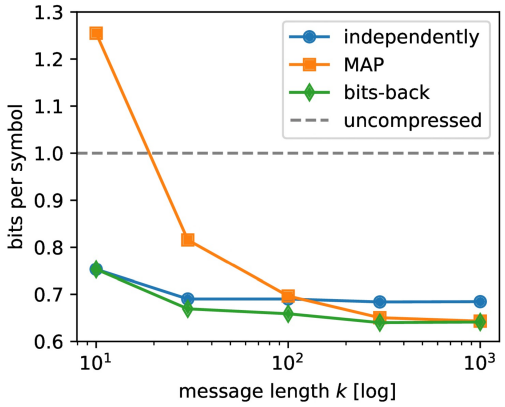
- $z \leftarrow$ decode from compressed using $P(\underline{Z})$
- $\underline{x} \leftarrow$ decode from compressed using $P(\underline{X} \mid \underline{Z}=z)$
- encode z onto compressed using $P(\underline{Z} \mid \underline{X}=\underline{x})$
- return ($\underline{x}$, compressed)

net bit rate of bits-back coding:

$$R_{net}(\underline{x}) = -\log P(\underline{X}=\underline{x} \mid \underline{z}=\underline{z}) - \log P(\underline{z}=\underline{z}) - (-\log P(\underline{z}=\underline{z} \mid \underline{X}=\underline{x}))$$

$$= -\log \frac{P(\underline{X}=\underline{x}, \underline{z}=\underline{z})\, P(\underline{X}=\underline{x})}{P(\underline{X}=\underline{x}, \underline{z}=\underline{z})} = -\log P(\underline{X}=\underline{x})$$

→ bits-back coding is optimal (net).



Next steps:

- How do we en/de-code fractional numbers of bits with stack semantics.
- What if we don't know the exact posterior?
  → Lecture 7: approximate Bayesian inference
- How can we efficiently train deep latent variable models?
  → Lecture 7 & subsequent: • variational expectation maximization
  • deep generative models