# CSC 412
# Probabilistic Learning and Reasoning
## Week 13 : Final Exam Review

Denys Linkov

University of Toronto

# Final exam logistic

- Final exam will be held in person on April 29, at 9am-12pm Toronto local time in room BN 322 (all sections).
- Exams will be 100 points in total and 180 mins long. Students are required to be at the exam location at least 10 mins early, with valid identification. Exam will be administered by FAS.
- You can use two optional A4 aid sheets - double-sided.
- Exam covers all lectures (weeks 1-12), it is closed book/internet.
- A representative practice exam will be posted on the webpage.

# Probabilistic ML Terminology

The final exam will be on the entire course; however, it will be more weighted towards post-midterm material. For pre-midterm material, refer to the midterm review slides on the website.

- Exponential families
- Directed Graphical Models
- Markov Random Fields
- Message passing
- Belief propagation
- Variable eliminitaion
- Sampling methods
- Markov chain Monte Carlo
- Variational Inference
- Variational Autoencoders
- Embeddings and Attention
- Constrained Decoding
- Speculative Decoding
- Diffusion models

We can write this distribution as an exponential family

$$\begin{aligned} p(x|\theta) &= \theta^x (1-\theta)^{1-x} \\ &= \exp\{x \log(\theta) + (1-x)\log(1-\theta)\} \\ &= \exp\{x \log(\tfrac{\theta}{1-\theta}) + \log(1-\theta)\} \end{aligned}$$

Here,

$$\begin{aligned} T(x) &= x \\ \eta &= \log(\tfrac{\theta}{1-\theta}) \\ A(\eta) &= \log(1+e^{\eta}) \\ h(x) &= 1 \end{aligned}$$

Notice that $A'(\eta) = \frac{e^{\eta}}{1+e^{\eta}} = \theta$ is the mean of $T(X) = X$ and $A''(\eta) = \frac{e^{\eta}}{(1+e^{\eta})^2} = \theta(1-\theta)$ is the variance of $X$.

# Mean of sufficient statistics

Moments of exponential families can be easily computed using the log-partition function. Let $X \sim p(x|\eta)$ and denote by $A'(\eta) = dA(\eta)/d\eta$

$$\begin{aligned}
\mathbb{E}[T(X)] - A'(\eta) &= \int T(x)p(x|\eta)dx - A'(\eta) \\
&= \int \{T(x) - A'(\eta)\}h(x)\exp\{\eta^\top T(x) - A(\eta)\}dx \\
&= \int \frac{d}{d\eta}\left(h(x)\exp\{\eta^\top T(x) - A(\eta)\}\right)dx \\
&= \frac{d}{d\eta}\int p(x|\eta)dx \\
&= \frac{d}{d\eta}1 = 0.
\end{aligned}$$

Thus, we conclude that $\mathbb{E}_\eta[T(X)] = A'(\eta)$.

The variance $\mathrm{var}_\eta(T(X))$ can be computed similarly.

# MLE for general Exponential Families

Recall: $p(x|\eta) = h(x)\exp\{\eta^\top T(x) - A(\eta)\}$.

After observing data $\mathcal{D}$ with $N$ samples, we write the log-likelihood:

$$\ell(\eta;\mathcal{D}) = \log p(\mathcal{D};\eta) = \sum_{i=1}^{N} \log h(x^{(i)}) + \eta^\top \sum_{i=1}^{N} T(x^{(i)}) - NA(\eta)$$

For the MLE derivation we solve:

$$\ell'(\eta;\mathcal{D}) = \sum_{i=1}^{N} T(x^{(i)}) - NA'(\eta) = 0$$

The MLE for the natural parameters $\eta$ of a general exponential family:

$$\hat{\eta}_{\text{MLE}} \quad \text{that solves} \quad A'(\hat{\eta}_{\text{MLE}}) = \frac{1}{N}\sum_{i=1}^{N} T(x^{(i)}).$$

Note: This equation may not have an explicit solution but the solution always corresponds to the global maximum

# Week 8-9: KL divergence

We will measure the difference between $q$ and $p$ using the **Kullback-Leibler divergence**

$$KL(q(z)||p(z)) = \int q(z) \log \frac{q(z)}{p(z)} dz$$

$$\text{or} = \sum_z q(z) \log \frac{q(z)}{p(z)}$$

Properties of the KL Divergence

- $KL(q||p) \geq 0$
- $KL(q||p) = 0 \Leftrightarrow q = p$
- $KL(q||p) \neq KL(p||q)$
- KL divergence is not a metric, since it's not symmetric

# A1 and Tutorial 1

- Review how to set constraints for exponential families
- Finding sufficient statistics

# Week 8-9: I & M Projection

- I-projection: $q^* = \arg\min_{q \in Q} KL(q||p) = \mathbb{E}_{x \sim q(x)} \log \frac{q(x)}{p(x)}$:
  - $p \approx q \implies KL(q||p)$ small
  - I-projection underestimates support, and does not yield the correct moments.
  - $KL(q||p)$ penalizes $q$ having mass where $p$ has none.
- M-projection: $q^* = \arg\min_{q \in Q} KL(p||q) = \mathbb{E}_{x \sim p(x)} \log \frac{p(x)}{q(x)}$:
  - $p \approx q \implies KL(p||q)$ small
  - $KL(p||q)$ penalizes $q$ missing mass where $p$ has some.
  - M-projection yields a distribution $q(x)$ with the correct mean and covariance.
- One way to proceed is the mean-field approach where we assume:

$$q(x) = \prod_{i \in V} q_i(x_i)$$

the set $Q$ is composed of those distributions that factor out.

# Week 9: Evidence Lower Bound

ELBO is a lower bound on the (log) evidence. Maximizing the ELBO is the same as minimizing $KL(q_\phi(z)||p(z|x))$.

$$
\begin{aligned}
KL(q_\phi(z)||p(z|x)) &= \mathop{\mathbb{E}}_{z\sim q_\phi} \log \frac{q_\phi(z)}{p(z|x)} \\
&= \mathop{\mathbb{E}}_{z\sim q_\phi}\left[ \log\left( q_\phi(z) \cdot \frac{p(x)}{p(z,x)} \right)\right] \\
&= \mathop{\mathbb{E}}_{z\sim q_\phi}\left[ \log \frac{q_\phi(z)}{p(z,x)} \right] + \mathop{\mathbb{E}}_{z\sim q_\phi} \log p(x) \\
&:= -\mathcal{L}(\phi) + \log p(x)
\end{aligned}
$$

Where $\mathcal{L}(\phi)$ is the **ELBO**:

$$
\mathcal{L}(\phi) = \mathop{\mathbb{E}}_{z\sim q_\phi}\left[ \log p(z,x) - \log q_\phi(z) \right]
$$

# Week 9: ELBO

- Rearranging, we get

$$\mathcal{L}(\phi) + KL(q_\phi(z)||p(z|x)) = \log p(x)$$

- Because $KL(q_\phi(z)||p(z|x)) \geq 0$,

$$\mathcal{L}(\phi) \leq \log p(x)$$

- maximizing the ELBO $\Rightarrow$ minimizing $KL(q_\phi(z)||p(z|x))$.

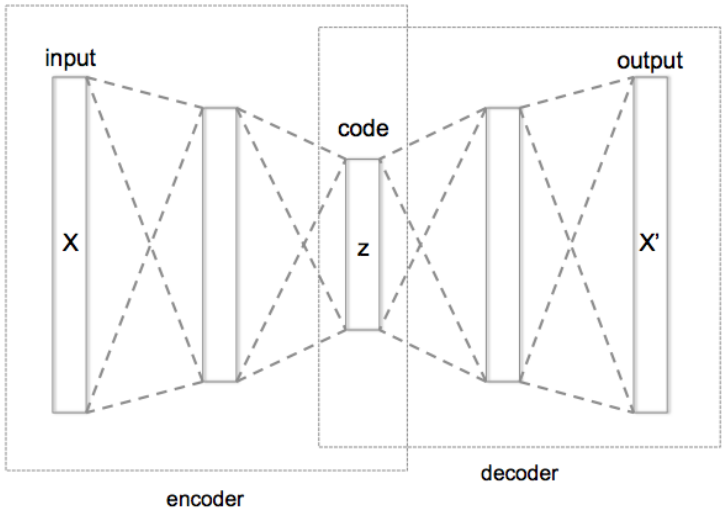
# Week 10: Autoencoders

Autoencoders reconstruct their input via an encoder and a decoder.

- **Encoder**: $g(x) = z \in F, \quad x \in X$
- **Decoder**: $f(z) = \tilde{x} \in X$
- where $X$ is the data space, and $F$ is the feature (latent) space.
- $z$ is the code, compressed representation of the input, $x$. It is important that this code is a bottleneck, i.e. that

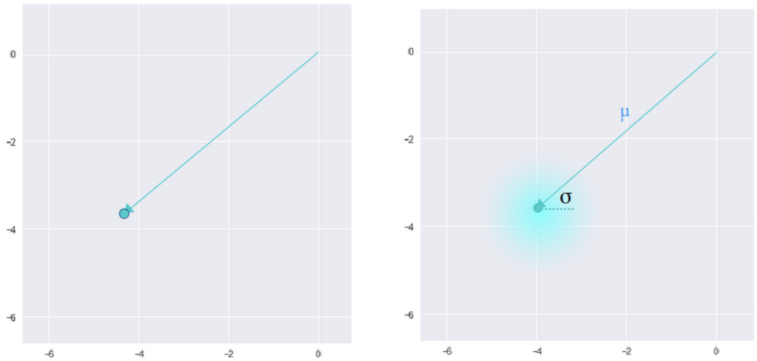
$$\dim F \ll \dim X$$

- Goal: $\tilde{x} = f(g(x)) \approx x$.

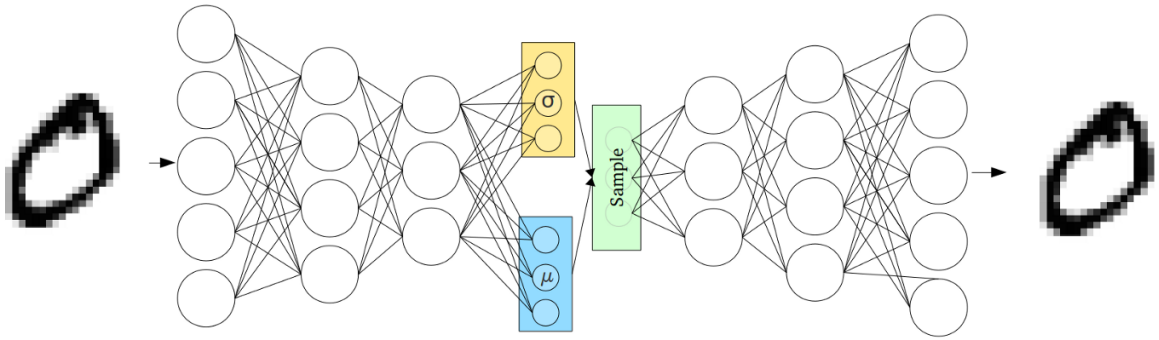


# Week 10: Variational Autoencoders

- The mean $\mu$ controls where encoding of input is centered while the standard deviation controls how much can the encoding vary.




- Encodings are generated at random from the "circle", the decoder learns that all nearby points refer to the same input.



# Week 10: VAE vs Amortized VAE Pipeline

- For a given input (or minibatch) $x_i$,

  - **Standard VAE**
  - Sample $z_i \sim q_{\phi_i}(z|x_i) = \mathcal{N}(\mu_i, \sigma_i^2 I)$.

  - **Amortized VAE**
  - Sample $z_i \sim q_\phi(z|x_i) = \mathcal{N}(\mu_\phi(x_i), \Sigma_\phi(x_i))$

- Run the code through decoder and get likelihood: $p_\theta(x|z)$.
- Compute the loss function (-ELBO):
  $L(x;\theta,\phi) = -E_{z_\phi \sim q_\phi}\Big[\log p_\theta(x|z)\Big] + KL(q_\phi(z|x)||p(z))$
- Use gradient-based optimization to backpropogate $\partial_\theta L$, $\partial_\phi L$

# LLM explainability

- SAE architecture - Figure out model features
- MoEs - Train with features in mind
- MoE model routing, sparse vs dense models
- Data selection for training

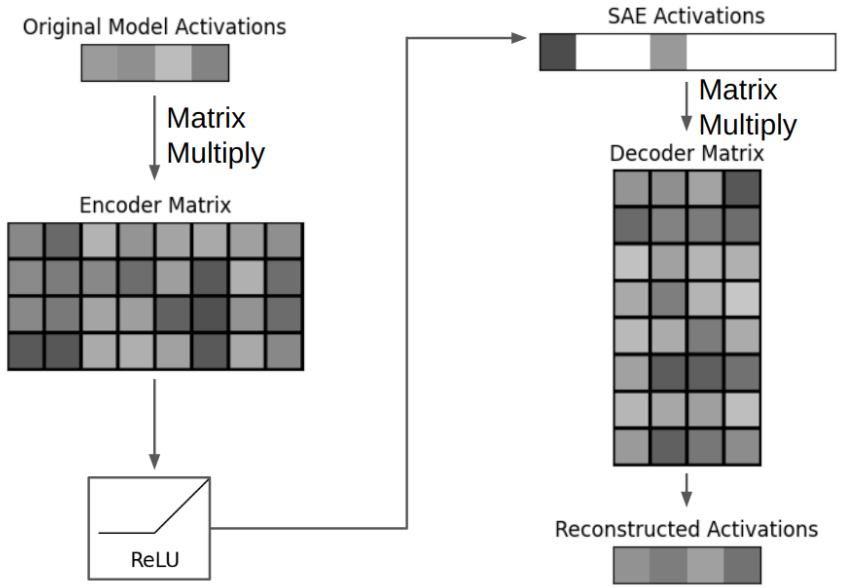



Figure: SAE architecture (Karvonen, 2024)

# Constrained decoding

- LLM sampling - Top k, Top p, Epsilon, Temperature
- Beam vs Greedy
- Approaches to limit output vocabulary
- Create practice problems to constrain

**Algorithm 1** Constrained Decoding

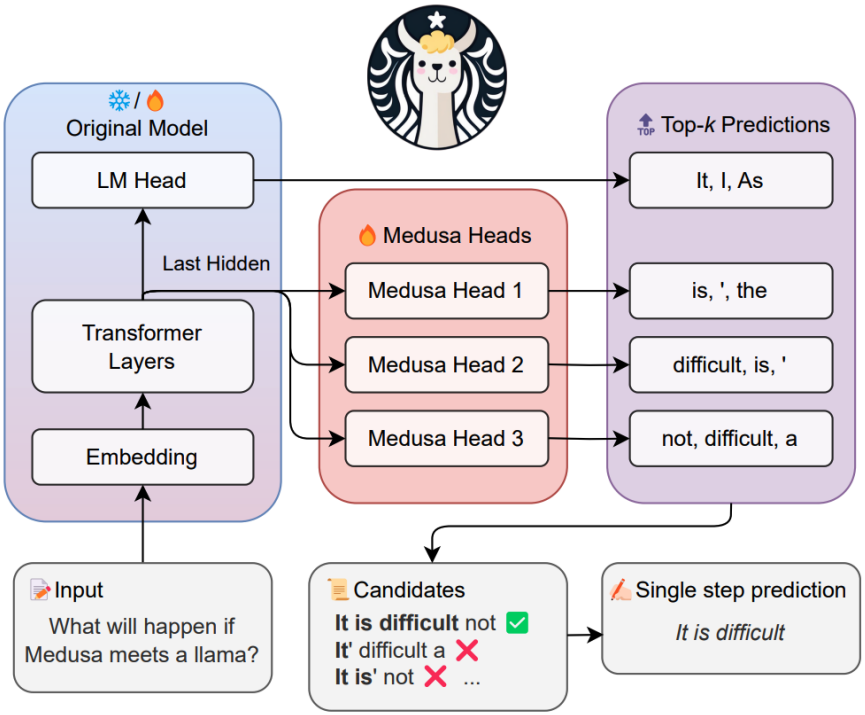
**Input:** Checker $C$, LLM $f$, Tokenized Prompt $x$
**Output:** Completion $o$ adhering to $C$

$o \leftarrow []$
$C.\texttt{init}()$
**loop**
  $C.\texttt{update}(o)$ // advance state of $C$
  $\boldsymbol{m} \leftarrow C.\texttt{mask}()$ // compute mask
  $\boldsymbol{v} \leftarrow f(x + o)$ // compute logits
  $\boldsymbol{v}' \leftarrow \boldsymbol{m} \odot \boldsymbol{v}'$
  $t \leftarrow \texttt{decode}(\alpha')$ // e.g., argmax or sample
  **if** $t = EOS$ **then break**
  $o.\texttt{append}(t)$
**end loop**
**return** $o$ // optionally detokenize

Figure: Constrained decoding formulation (Beurer-Kellner, 2024)

# Speculative decoding

- Approaches - Draft (small model) and verify (large model)
- Medusa head approach, predict $k$ additional Tokens
- Review empirical vs theoretical results




Medusa

**Algorithm 1** SpeculativeDecodingStep

**Inputs:** $M_p, M_q, prefix$.
▷ Sample $\gamma$ guesses $x_{1,\ldots,\gamma}$ from $M_q$ autoregressively.
**for** $i = 1$ **to** $\gamma$ **do**
  $q_i(x) \leftarrow M_q(prefix + [x_1, \ldots, x_{i-1}])$
  $x_i \sim q_i(x)$
**end for**
▷ Run $M_p$ in parallel.
$p_1(x), \ldots, p_{\gamma+1}(x) \leftarrow$
  $M_p(prefix), \ldots, M_p(prefix + [x_1, \ldots, x_\gamma])$
▷ Determine the number of accepted guesses $n$.
$r_1 \sim U(0,1), \ldots, r_\gamma \sim U(0,1)$
$n \leftarrow \min(\{i - 1 \mid 1 \le i \le \gamma, r_i > \frac{p_i(x)}{q_i(x)}\} \cup \{\gamma\})$
▷ Adjust the distribution from $M_p$ if needed.
$p'(x) \leftarrow p_{n+1}(x)$
**if** $n < \gamma$ **then**
  $p'(x) \leftarrow norm(max(0, p_{n+1}(x) - q_{n+1}(x)))$
**end if**
▷ Return one token from $M_p$, and $n$ tokens from $M_q$.
$t \sim p'(x)$
**return** $prefix + [x_1, \ldots, x_n, t]$

Speculative Decoding

# Closing remarks

Continuing with machine learning:

- Courses
  - CSC 413/2516, “Neural Networks and Deep Learning”
  - CSC 2515, “Machine Learning”
  - CSC 2532, “Statistical Learning Theory”
  - CSC 2541, ”Neural Network Training Dynamics”
  - Topics courses (varies from year to year): Reinforcement Learning, Algorithmic Fairness, Computer Vision w/ ML, NLP w/ ML, Health w/ ML etc.
  - CSC49X - Capstone courses
- Videos from top ML conferences (NeurIPS, ICML, ICLR)
- Try to reproduce results from papers
  - If they’ve released code, you can use that as a guide if you get stuck.
- Lots of excellent free resources available online!

# Summary

- Review lectures.
- Understand **derivations**.
- Solve the practice final.
- Review papers mentioned in class
- Fill out course evaluations!
- Thanks!