# CS-GY 6923: Lecture 4
# Continue on Bayesian Perspective, Modeling Language

NYU Tandon School of Engineering, Akbar Rafiey

# Course news

- First written problem set due next Wednesday Feb. 19.
  - I will hold a Zoom office hour for the homework.
  - We will release solutions and go over them in office hours.
- Lab 3 was released yesterday, due on Feb 26.

# Recap: probabilistic modeling

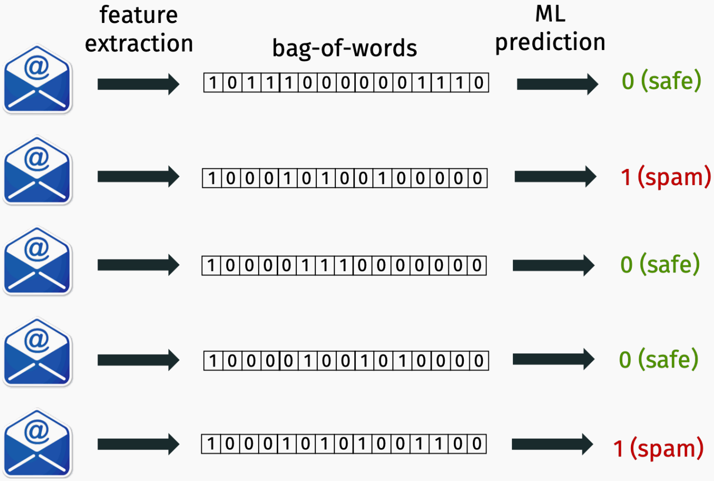
In a Bayesian or Probabilistic approach to machine learning we always start by conjecturing a

**probabilistic model**

that plausibly could have generated our data.

- The model guides how we make predictions.
- The model typically has unknown parameters $\vec{\theta}$ and we try to find the most reasonable parameters based on observed data.

# Recap: spam prediction



feature extraction

bag-of-words

ML prediction

1 0 1 1 1 0 0 0 0 0 0 1 1 1 0 → 0 (safe)

1 0 0 0 1 0 1 0 0 1 0 0 0 0 0 → 1 (spam)

1 0 0 0 0 1 1 1 0 0 0 0 0 0 0 → 0 (safe)

1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 → 0 (safe)

1 0 0 0 1 0 1 0 1 0 0 1 1 0 0 → 1 (spam)

# Recap: email model

Include each word in an email with some fixed probability. That probability will differ depending on whether or not it is a spam or regular email.

| **Not Spam** | **Spam** |
|---|---|
| $p_{\text{won},0} = .02$ | $p_{\text{won},1} = .1$ |
| $p_{\$,0} = .05$ | $p_{\$,1} = .2$ |
| $p_{student,0} = .06$ | $p_{student,1} = .01$ |

# Recap: probabilistic model for email

**Probabilistic model** for (bag-of-words, label) pair $(\mathbf{x}, y)$:

- Set $y = 0$ (not spam) with probability $p_0$ and $y = 1$ (spam) with probability $p_1 = 1 - p_0$.
  - $p_0$ is probability an email is not spam (e.g. 99%).
  - $p_1$ is probability an email is spam (e.g. 1%).
- If $y = 0$, for each $i$, set $x_i = 1$ with prob. $p_{i0}$.
- If $y = 1$, for each $i$, set $x_i = 1$ with prob. $p_{i1}$.

**Unknown model parameters:**

- $p_0, p_1$,
- $p_{10}, p_{20}, \ldots p_{d0}$, one for each of the $d$ vocabulary words.
- $p_{11}, p_{21}, \ldots p_{d1}$, one for each of the $d$ vocabulary words.

# Recap: parameter estimation

**Reasonable way to set parameters:**

- Set $p_0$ and $p_1$ to the empirical fraction of not spam/spam emails.
- For each word $i$, set $p_{i0}$ to the empirical probability word $i$ appears in a $\underline{\text{non-spam}}$ email.
- For each word $i$, set $p_{i1}$ to the empirical probability word $i$ appears in a $\underline{\text{spam}}$ email.

# Done with modeling

## on to prediction

# Probability review

- **Probability:** p(x) – the probability event $x$ happens.
- **Joint probability:** p(x,y) – the probability that event $x$ and event $y$ happen.
- **Conditional Probability** $p(x \mid y)$ – the probability $x$ happens given that $y$ happens.

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

# Classification rule

Given unlabeled input $(\mathbf{w}, \_\_\_)$, choose the label $y \in \{0, 1\}$ which is most likely given the data. Recall $\mathbf{w} = [0, 0, 1, \ldots, 1, 0]$.

**Classification rule: maximum a posterior (MAP) estimate.**

**Step 1. Compute:**

- $p(y = 0 \mid \mathbf{w})$: prob. $y = 0$ given observed data vector $\mathbf{w}$.
- $p(y = 1 \mid \mathbf{w})$: prob. $y = 1$ given observed data vector $\mathbf{w}$.

**Step 2. Output:** 0 or 1 depending on which probability is larger.

$p(y = 0 \mid \mathbf{w})$ and $p(y = 1 \mid \mathbf{w})$ are called **posterior** probabilities.

# Evaluating the posterior

How to compute the posterior? **Bayes rule!**

$$p(y = 0 \mid \mathbf{w}) = \frac{p(\mathbf{w} \mid y = 0)p(y = 0)}{p(\mathbf{w})} \tag{1}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{2}$$

- **Prior:** Probability in class 0 prior to seeing any data.
- **Posterior:** Probability in class 0 after seeing the data.

# Evaluating the posterior

Goal is to determine which is larger:

$$p(y=0 \mid \mathbf{w}) = \frac{p(\mathbf{w} \mid y=0)p(y=0)}{p(\mathbf{w})} \qquad \text{vs.}$$

$$p(y=1 \mid \mathbf{w}) = \frac{p(\mathbf{w} \mid y=1)p(y=1)}{p(\mathbf{w})}$$

- We can ignore the evidence $p(\mathbf{w})$ since it is the same for both sides!
- $p(y=0)$ and $p(y=1)$ already known (computed from training data). These are our computed parameters $p_0$, $p_1$.
- $p(\mathbf{w} \mid y=0) = ?$ $p(\mathbf{w} \mid y=1) = ?$

# Evaluating the posterior

Consider the example $\mathbf{w} = [0, 1, 1, 0, 0, 0, 1, 0]$.

Recall that, under our model, index $i$ is 1 with probability $p_{i0}$ if we are not spam, and 1 with probability $p_{i1}$ if we are spam .

$$p(\mathbf{w} \mid y = 0) =$$

$$p(\mathbf{w} \mid y = 1) =$$

# Naive bayes classifier

**Training/Modeling:** Use existing data to compute:

- $p_0 = p(y = 0), p_1 = p(y = 1)$
- For all $i$ compute:
  - $p_{i0} = p(w_i = 1 \mid y = 0)$ and $(1 - p_{i0}) = p(w_i = 0 \mid y = 0)$
  - $p_{i1} = p(w_i = 1 \mid y = 1)$ and $(1 - p_{i1}) = p(w_i = 0 \mid y = 1)$

**Prediction:**

- For new input $\mathbf{w}$:
  - Compute $p(\mathbf{w} \mid y = 0) = \prod_i p(w_i \mid y = 0)$
  - Compute $p(\mathbf{w} \mid y = 1) = \prod_i p(w_i \mid y = 1)$
- Return $y = 0$ if
  $p\left(\mathbf{w} \mid y = 0\right) \cdot p\left(y = 0\right) \geq p\left(\mathbf{w} \mid y = 1\right) \cdot p\left(y = 1\right)$
- Return $y = 1$ otherwise

# Other applications of the bayesian perspective on regression

# Bayesian regression

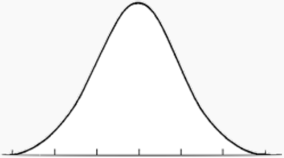
The Bayesian view offers an interesting alternative perspective on many machine learning techniques.

**Example:** Linear Regression.

**Probabilistic model:**

$$y = \langle \mathbf{x}, \boldsymbol{\beta} \rangle + \eta$$

where the $\eta$ drawn from $N(0, \sigma^2)$ is **random Gaussian noise**.



$$Pr(\eta = z) \sim$$

The symbol $\sim$ means "is proportional to".

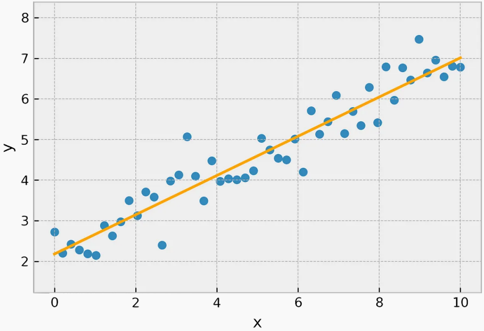
# Illustration

**Example:** Linear Regression.

**Probabilistic model:**

$$y = \langle \mathbf{x}, \boldsymbol{\beta} \rangle + \eta$$

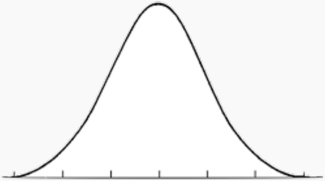
where the $\eta$ drawn from $N(0, \sigma^2)$ is **random Gaussian noise**. The noise is independent for different inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$.



# Gaussian distribution refresher

**Names for same thing:** Normal distribution, Gaussian distribution, bell curve.

Parameterized by mean $\mu$ and variance $\sigma^2$.

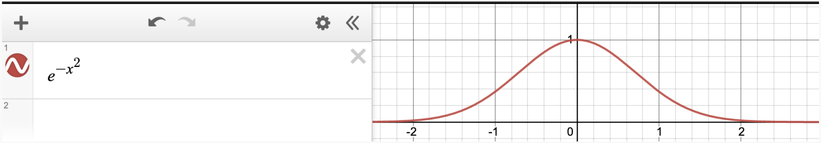


$\eta$ is a continuous random variable, so it has a probability density function $p(\eta)$ with $\int_{-\infty}^{\infty} p(\eta) d\eta = 1$

$$p(\eta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{\eta-\mu}{\sigma})^2}$$

# Gaussian distribution refresher

The important thing to remember is that the the PDF falls off exponentially as we move further from the mean.



$e^{-x^2}$

The normalizing constant in front 1/2, etc. don't matter so much.

# Bayesian regression

## How should we select $\beta$ for our model?

Also use a Bayesian approach!

**First thought**: choose $\beta$ to maximize:

$$\text{posterior} = \Pr(\beta \mid \mathbf{X}, \mathbf{y}) = \frac{\Pr(\mathbf{X}, \mathbf{y} \mid \beta) \Pr(\beta)}{\Pr(\mathbf{X}, \mathbf{y})} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}.$$

But in this case, we don't have a prior – no values of $\beta$ are inherently more likely than others.

Choose $\beta$ to maximize just the likelihood:

$$\frac{\Pr(\mathbf{X}, \mathbf{y} \mid \beta)\cancel{\Pr(\beta)}}{\cancel{\Pr(\mathbf{X}, \mathbf{y})}} = \frac{\text{likelihood} \times \cancel{\text{prior}}}{\cancel{\text{evidence}}}.$$

This is called the **maximum likelihood estimate**.

# Fixed design linear regression

- Often we think of $\mathbf{X}$ as fixed and deterministic, and only $\mathbf{y}$ is generated at random in the model. This is called the fixed design setting.
- We can also consider a randomized design setting, but it is slightly more complicated.
- In the fixed design setting our task of maximizing $\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta})$ simplifies to maximizing

$$\max_{\boldsymbol{\beta}} \Pr(\mathbf{y} \mid \boldsymbol{\beta})$$

# Maximum likelihood estimate

**Data:**

$$\mathbf{X} = \begin{bmatrix} — & \mathbf{x}_1 & — \\ — & \mathbf{x}_2 & — \\ & \vdots & \\ — & \mathbf{x}_n & — \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

**Model:** $y_i = \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle + \eta_i$ where $p(\eta_i = z) \sim e^{-z^2/2\sigma^2}$ and $\eta_1, \ldots, \eta_n$ are independent.

Exercise:

$$\Pr(\mathbf{y} \mid \boldsymbol{\beta}) \sim$$

# Log likelihood

Easier to work with the **log likelihood**:

$$\begin{aligned}
\arg\max_{\boldsymbol{\beta}} \Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) &= \arg\max_{\boldsymbol{\beta}} \prod_{i=1}^{n} e^{-(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 / 2\sigma^2} \\
&= \arg\max_{\boldsymbol{\beta}} \ \log \left( \prod_{i=1}^{n} e^{-(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 / 2\sigma^2} \right) \\
&= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} -(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2 / 2\sigma^2 \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta} \rangle)^2.
\end{aligned}$$

# Maximum likelihood estimator

**Conclusion:** Choose $\beta$ to minimize:

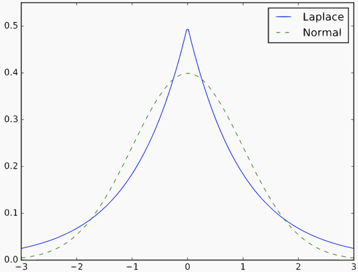
$$\sum_{i=1}^{n}(y_i - \langle \mathbf{x}_i, \beta\rangle)^2 = \|\mathbf{y} - \mathbf{X}\beta\|_2^2.$$

**This is a completely different justification for squared loss!**

Minimizing the $\ell_2$ loss is "optimal" when you assume your data follows a linear model with i.i.d. Gaussian noise.

# Bayesian regression

If we had modeled our noise $\eta$ as Laplace noise, we would have found that minimizing $\|\mathbf{y} - \mathbf{X}\beta\|_1$ was optimal.



$$Pr(\eta = z) \sim$$

Laplace noise has “heavier tails”, meaning that it results in more outliers.

**This is a completely different justification for $\ell_1$ loss.**

# Bayesian regularization

- We can add another layer of probabilistic modeling by also assuming $\boldsymbol{\beta}$ is random and comes from some distribution, which encodes our prior belief on what the parameters are.
- Recall Maximum a posteriori (MAP estimation):

$$\Pr(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) = \frac{\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \Pr(\boldsymbol{\beta})}{\Pr(\mathbf{X}, \mathbf{y})}.$$

- Assume values in $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_d]$ come from some distribution.
- What is a common model?

# Bayesian regularization

- **Common model:** Each $\beta_i$ drawn from $N(0, \gamma^2)$, i.e. normally distributed, independent.
- Encodes a belief that we are unlikely to see models with very large coefficients.

**Goal:** choose $\boldsymbol{\beta}$ to maximize:

$$\Pr(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) = \frac{\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \Pr(\boldsymbol{\beta})}{\Pr(\mathbf{X}, \mathbf{y})}.$$

# Bayesian regularization

**Goal:** choose $\boldsymbol{\beta}$ to maximize:

$$\Pr(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) = \frac{\Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \Pr(\boldsymbol{\beta})}{\Pr(\mathbf{X}, \mathbf{y})}.$$

- We can still ignore the "evidence" term $\Pr(\mathbf{X}, \mathbf{y})$ since it is a constant that does not depend on $\boldsymbol{\beta}$.
- $\Pr(\boldsymbol{\beta}) = \Pr(\beta_1) \cdot \Pr(\beta_2) \cdot \ldots \cdot \Pr(\beta_d)$
- If each $\beta_i$ drawn from $N(0, \gamma^2)$, $\Pr(\boldsymbol{\beta}) \sim ?$

# Bayesian regularization

Easier to work with the **log likelihood**:

$$
\begin{aligned}
&\arg\max_{\boldsymbol{\beta}} \Pr(\mathbf{X}, \mathbf{y} \mid \boldsymbol{\beta}) \cdot \Pr(\boldsymbol{\beta}) \\
&= \arg\max_{\boldsymbol{\beta}} \prod_{i=1}^{n} e^{-(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle)^2/2\sigma^2} \cdot \prod_{i=1}^{n} e^{-(\beta_i)^2/2\gamma^2} \\
&= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} -(y_i - \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle)^2/2\sigma^2 + \sum_{i=1}^{d} -(\beta_i)^2/2\gamma^2 \\
&= \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \langle \mathbf{x}_i, \boldsymbol{\beta}\rangle)^2 + \frac{\sigma^2}{\gamma^2} \sum_{i=1}^{d} (\beta_i)^2
\end{aligned}
$$

Choose $\boldsymbol{\beta}$ to minimize $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\sigma^2}{\gamma^2}\|\boldsymbol{\beta}\|_2^2$.

**Completely different justification for ridge regularization!**

# Bayesian regularization

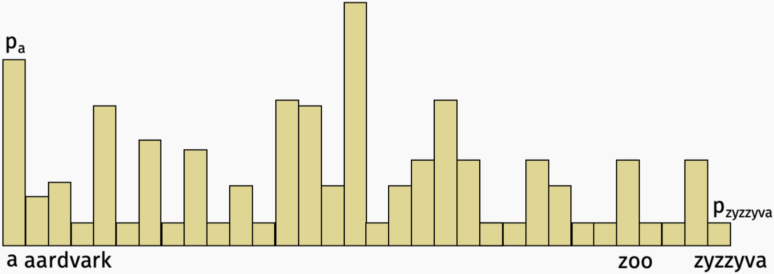
**Test your intuition:** What modeling assumption justifies LASSO regularization: $\min \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$?

# Modeling language

# Generative ML

**Key idea behind generative ML:**

- Build a very good probabilistic model for your data.
- Use that model to generate realistic looking new data.



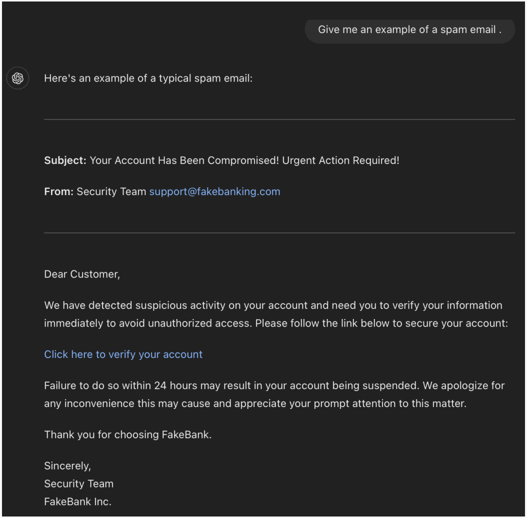
# Example

**Key idea behind generative ML:** Build a very good probabilistic model for your data. Use that model to generate realistic looking new data.

**Email example from our model:** *Keeps retaining in astro associated to no garden superconducting whistleblower on effusion eigenvalue jobs worker for car shortlist villa depictions fitness the easter veto devices expressed secondary user metal this administrative the do of to struct coffee online cde the open through requirement stamps you job g thus drop stations.*

# Language modeling

How do we go from this to something more like what modern models can produce?



Give me an example of a spam email .

Here's an example of a typical spam email:

**Subject:** Your Account Has Been Compromised! Urgent Action Required!

**From:** Security Team support@fakebanking.com

Dear Customer,

We have detected suspicious activity on your account and need you to verify your information immediately to avoid unauthorized access. Please follow the link below to secure your account:

Click here to verify your account

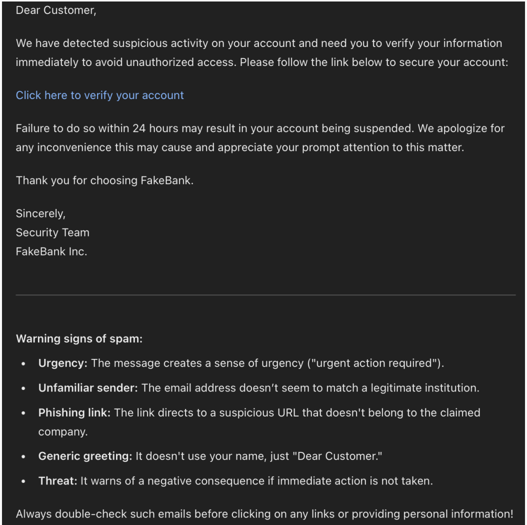
Failure to do so within 24 hours may result in your account being suspended. We apologize for any inconvenience this may cause and appreciate your prompt attention to this matter.

Thank you for choosing FakeBank.

Sincerely,
Security Team
FakeBank Inc.

# Language modeling

How do we go from jumbled words to something more like what modern models can produce?



Dear Customer,

We have detected suspicious activity on your account and need you to verify your information immediately to avoid unauthorized access. Please follow the link below to secure your account:

Click here to verify your account

Failure to do so within 24 hours may result in your account being suspended. We apologize for any inconvenience this may cause and appreciate your prompt attention to this matter.

Thank you for choosing FakeBank.

Sincerely,
Security Team
FakeBank Inc.

---

**Warning signs of spam:**

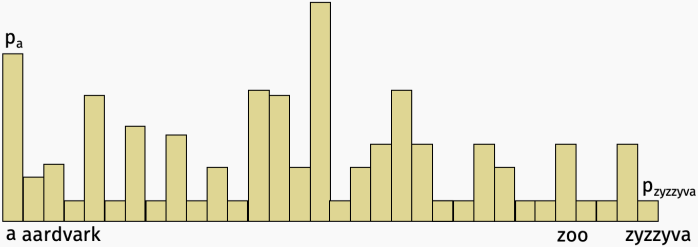
- **Urgency:** The message creates a sense of urgency ("urgent action required").
- **Unfamiliar sender:** The email address doesn’t seem to match a legitimate institution.
- **Phishing link:** The link directs to a suspicious URL that doesn't belong to the claimed company.
- **Generic greeting:** It doesn't use your name, just "Dear Customer."
- **Threat:** It warns of a negative consequence if immediate action is not taken.

Always double-check such emails before clicking on any links or providing personal information!

# Autogressive models

**Main issue:** Our model lacks context!

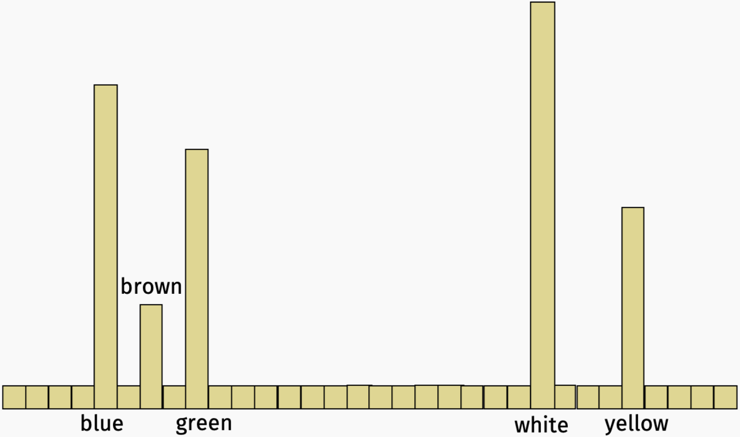
The color of the dress is ______.



# Autogressive models

**Main issue:** Our model lacks context!

The color of the dress is ______.



# Autogressive models

**Key idea:** Distribution that a word is chosen from should depend on previous words in the sentence/paragraph.

Consider generating a sentence with words $x_1, x_2, \ldots, x_n$.

- Initialize the first word $x_1$ of the sentence (e.g. at random).
- Choose $x_2$ based on $x_1$.
- Choose $x_3$ based on $x_1, x_2, \ldots$

Concretely, set $x_i = w$ with probability:

$$P(x_i = w \mid x_{i-1}, x_{i-2}, \ldots, x_1).$$

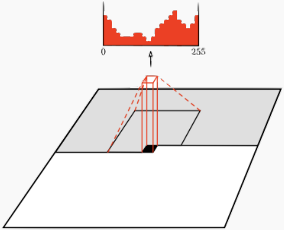
# Autogressive models

Autoregressive model's generate text in order.

- How most humans write sentences, emails, short text.
- How the latest modern language models write text (e.g. the GPT family of models.)

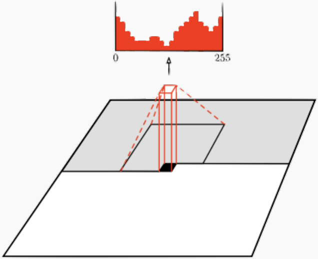
**This is not the only approach to generative modeling, but it is one that works fairly well in practice, especially for text.**

Can also be used e.g. for images, but no longer state of the art.



# Autogressive models

Can also be used e.g. for images, but no longer state of the art.



$$p(x_i) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \dots p(x_i \mid x_1, x_2, \dots, x_{i-1})$$

$$p(x_i, \theta) = \prod_{j=1}^{i} p(x_j \mid x_1, \dots, x_{j-1}; \theta) \quad \text{(parametrize a model by } \theta\text{)}$$

# Autogressive models

Can also be used e.g. for images, but no longer state of the art.

$$p(x_i) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \dots p(x_i \mid x_1, x_2, \dots, x_{i-1})$$

$$p(x_i, \theta) = \prod_{j=1}^{i} p(x_j \mid x_1, \dots, x_{j-1}; \theta) \quad \text{(parametrize a model by } \theta\text{)}$$

$$\log p(x_i, \theta) = \sum_{j=1}^{i} \log p(x_j \mid x_1, \dots, x_{j-1}; \theta) \qquad \text{(MLE)}$$

Different architectures to model $p(x_j \mid x_1, \dots, x_{j-1}; \theta)$:

- Suggestion?

# Autogressive models

Different architectures to model $p(x_j \mid x_1, \ldots, x_{j-1}; \theta)$:

- Convolutional Models (PixelCNN)
- Recurrent Models (PixelRNN)
- Transformer-based Models (ImageGPT, GPT)
- All of the above approaches need good (semantic) word embeddings. We will learn about this later in the course.
- Today: simple model and simple complexity

# Limited lookback

**Key idea:** Distribution that a word is chosen from should depend on previous **k words** in the sentence/paragraph. $k$ is a parameter that controls model complexity.

# Limited lookback

**Key idea:** Distribution that a word is chosen from should depend on previous **k words** in the sentence/paragraph. $k$ is a parameter that controls model complexity.

Consider generating a sentence with words $x_1, x_2, \ldots, x_n$.

- Initialize the first $k$ word $x_1, \ldots, x_k$ of the sentence (e.g. at random).
- Choose $x_{k+1}$ based on $x_1, \ldots, x_k$.
- Choose $x_{k+2}$ based on $x_2, \ldots, x_{k+1}$.
- Choose $x_{k+3}$ based on $x_3, \ldots, x_{k+2}$.
- $\ldots$

Set $x_i = w$ with probability:

$$P(x_i = w \mid x_{i-1}, x_{i-2}, \ldots, x_{i-k}).$$

# Probabilistic model

Set $x_i = w$ with probability:

$$P(x_i = w \mid x_{i-k}, \dots, x_{i-2}, x_{i-1}).$$

This probability can be tractably estimate from our data!

It is exactly the same as the probability of observing the $k+1$-gram $[x_{i-k}, \dots, x_{i-2}, x_{i-1}, w]$.

**Training:**

- For corpus of text, collect all $k+1$-grams and record their frequency. (not much optimization involved!)

**Prediction:**

- At step $i$, sample from the subset of $k+1$ grams starting with $[x_{i-k}, \dots, x_{i-2}, x_{i-1}]$, with probability proportional to their frequency.

# Example

The color of the dress is ______.

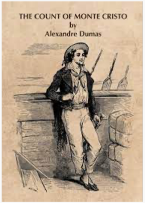
- Reasonable completions for $k = 2$:
- Reasonable completions for $k = 5$:

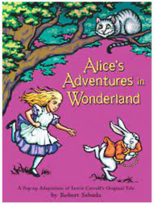
# Lab 3

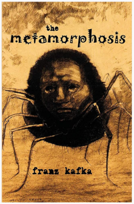
Credit to Raphael Meyer who was a Ph.D. student at CSE Tandon, currently a postdoc at Caltech.

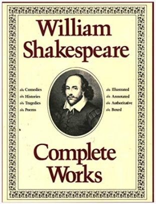
# Lab 3

- Train model on free books from Project Gutenberg.




Alice's Adventures in Wonderland
A Pop-up Adaptation of Lewis Carroll's Original Tale
by Robert Sabuda



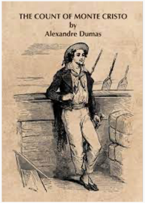


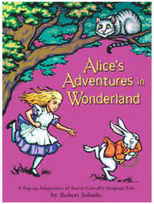
- Evaluate effect of changing $k$. Tradeoff between better performance and more “copying” from the course text.

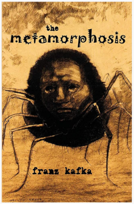
*Virtue itself of vice must pardon beg, Yea, curb and woo for leave to do him good, She shall undo her credit with the judge, or own great place, Could fetch your brother from the angry law; do no stain to your own souls so blind That you will clear yourself from all suspense.*

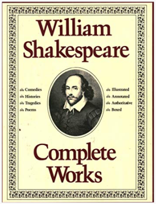
# Lab 3

- Train model on free books from Project Gutenberg.




Alice's Adventures in Wonderland
A Pop-up Adaptation of Lewis Carroll's Original Tale
by Robert Sabuda





- Evaluate effect of changing $k$. Tradeoff between better performance and more "copying" from the source text.

*During this time, Madame Morrel had told her all,—'Giovanni,' said she, 'you should have brought this child with you; we would have replaced the parents it has lost, have called it Benedetto, and then, in a loyal duel, and not in Arabia, and in France eternal friendships are as rare as the custom of doing when saying "Yes." "Good; he accepts," said Monte Cristo.*

Next time: How to deal with non-binary cases?