Log-Linear Tutorial
Examples on random variables

The gradient of the log linear model

The exponential family

Interactive visualization
Examples on Random Variables
What is a random variable?

- **A Random Variable** $X$ is a function that maps outcomes of random experiments to a set of properties. (We will dig more into this later.)
- **A Probability Distribution** $p(X=x)$ is a function that measures the probability that outcomes with the particular property $x$ will occur.

Example: rolling two dices.

Outcome: \{2, 3\}

Property: (1) sum of the numbers; (2) product of the numbers; ...

Value/measure: (1) 5; (2) 6; ...

---

adapted from A. Aldo Faisal, Cheng Soon Ong, and Marc Peter Deisenroth Mathematics for Machine Learning
Why do we need Random Variables?

- Random variables are fundamentally about interactions between different properties of elements of the sample space.
- Independence and correlation are properties of random variables and not of the probability spaces.

Example: rolling two dice.

1. The probability that (the sum is smaller than 5) while (the product is larger than 5)
2. ...
The Gradient of the Log-Linear Model
The Gradient of a Log-Linear Model

Finding the partial derivative

\[
L(\theta) = - \sum_{n=1}^{N} \log p(y_n | x_n, \theta)
\]

\[
\Pr(y | x) = \frac{\exp \left( \theta \cdot f(x, y) \right)}{\sum_{y' \in Y} \exp \left( \theta \cdot f(x, y') \right)}
\]

adapted from Noah A. Smith: https://homes.cs.washington.edu/~nasmith/papers/smith.tut04.pdf
The Gradient of a Log-Linear Model

\[ \sum_{j=1}^{m} \log \frac{\exp (\vec{\theta} \cdot \vec{f}(x, y))}{\sum_{y'} \exp (\vec{\theta} \cdot \vec{f}(x, y'))} \]

\[ \mathcal{L} (\vec{\theta}) = \left( \sum_{j=1}^{m} \vec{\theta} \cdot \vec{f}(x_j, y_j^*) \right) - \sum_{j=1}^{m} \log \sum_{y'} \exp \left( \vec{\theta} \cdot \vec{f}(x_j, y') \right) \]

adapted from Noah A. Smith: https://homes.cs.washington.edu/~nasmith/papers/smith.tut04.pdf
The Gradient of a Log-Linear Model

\[
\sum_{j=1}^{m} \log \frac{\exp (\vec{\theta} \cdot \vec{f}(x, y))}{\sum_{y'} \exp (\vec{\theta} \cdot \vec{f}(x, y'))}
\]

\[
\mathcal{L} (\vec{\theta}) = \left( \sum_{j=1}^{m} \vec{\theta} \cdot \vec{f}(x_j, y_j^*) \right) - \sum_{j=1}^{m} \log \sum_{y'} \exp (\vec{\theta} \cdot \vec{f}(x_j, y'))
\]

adapted from Noah A. Smith: https://homes.cs.washington.edu/~nasmith/papers/smith.tut04.pdf
The Gradient of a Log-Linear Model

\[ \mathcal{L}(\vec{\theta}) = \left( \sum_{j=1}^{m} \vec{\theta} \cdot \vec{f}(x_j, y_j^*) \right) - \sum_{j=1}^{m} \log \sum_{y'} \exp \left( \vec{\theta} \cdot \vec{f}(x_j, y') \right) \]

\[ \left( \sum_{j=1}^{m} \sum_{k} \theta_k f_k(x_j, y_j^*) \right) - \sum_{j=1}^{m} \log \sum_{y'} \exp \left( \sum_{k} \theta_k f_k(x_j, y') \right) \]

adapted from Noah A. Smith: https://homes.cs.washington.edu/~nasmith/papers/smith.tut04.pdf
The Gradient of a Log-Linear Model

\[ \mathcal{L} (\theta) = \left( \sum_{j=1}^{m} \sum_{k} \theta_k f_k(x_j, y^*_j) \right) - \sum_{j=1}^{m} \log \sum_{y'} \exp \left( \sum_{k} \theta_k f_k(x_j, y') \right) \]

Take the derivative with respect to \( \theta_\ell \), we have

\[ \frac{\partial \mathcal{L}}{\partial \theta_\ell} = \left( \sum_{j=1}^{m} f_\ell(x_j, y^*_j) \right) - \sum_{j=1}^{m} \frac{\sum_{y'} (\exp \sum_{k} \theta_k f_k(x_j, y')) f_\ell(x_j, y')}{\sum_{y'} \exp \sum_{k} \theta_k f_k(x_j, y')} \]

adapted from Noah A. Smith: https://homes.cs.washington.edu/~nasmith/papers/smith.tut04.pdf
The Gradient of a Log-Linear Model

\[ \frac{\partial L}{\partial \theta} \]

\[ \sum_{j=1}^{m} \log \left( \sum_{y'} \exp \left( \sum_{k} \theta_k f_k(x_j, y') \right) \right) \]

\( (\ln x)' = \frac{1}{x} \sum_{j=1}^{m} \frac{\sum_{y'} \exp \sum_{k} \theta_k f_k(x_j, y')}{\sum_{y'} \exp \sum_{k} \theta_k f_k(x_j, y')} \)

adapted from Noah A. Smith: https://homes.cs.washington.edu/~nasmith/papers/smith.tut04.pdf
The Gradient of a Log-Linear Model

\[ \frac{\partial L}{\partial \theta_{\ell}} \]

\[ \sum_{j=1}^{m} \log \sum_{y'} \exp \left( \sum_{k} \theta_k f_k(x_j, y') \right) \]

\[ (e^x)' = e^x \]

\[ \sum_{j=1}^{m} \frac{\sum_{y'} (\exp \sum_{k} \theta_k f_k(x_j, y'))}{\sum_{y'} \exp \sum_{k} \theta_k f_k(x_j, y')} \]

adapted from Noah A. Smith: https://homes.cs.washington.edu/~nasmith/papers smith.tut04.pdf
The Gradient of a Log-Linear Model

\[
\frac{\partial L}{\partial \theta_l} = \sum_{j=1}^{m} \log \sum_{y'} \exp \left( \sum_k \theta_k f_k(x_j, y') \right)
\]

\[
= \sum_{j=1}^{m} \frac{\sum_{y'} (\exp \sum_k \theta_k f_k(x_j, y')) f_l(x_j, y')}{\sum_{y'} \exp \sum_k \theta_k f_k(x_j, y')}
\]
The Gradient of a Log-Linear Model

\[
\left( \sum_{j=1}^{m} f_{\ell}(x_j, y_j^*) \right) - \sum_{j=1}^{m} \frac{\sum_{y'} \left( \exp \sum_k \theta_k f_k(x_j, y') \right) f_{\ell}(x_j, y')}{\sum_{y'} \exp \sum_k \theta_k f_k(x_j, y')}
\]

\[
\sum_{j=1}^{m} \left( f_{\ell}(x_j, y_j^*) - \frac{\sum_{y'} \left( \exp \sum_k \theta_k f_k(x_j, y') \right) f_{\ell}(x_j, y')}{\sum_{y'} \exp \sum_k \theta_k f_k(x_j, y')} \right)
\]

adapted from Noah A. Smith: https://homes.cs.washington.edu/~nasmith/papers/smith.tut04.pdf
The Gradient of a Log-Linear Model

\[
\sum_{j=1}^{m} \left( f_{\ell}(x_j, y_j^*) - \frac{\sum_{y'} \left( \exp \sum_{k} \theta_k f_k(x_j, y') \right) f_{\ell}(x_j, y')} {\sum_{y'} \exp \sum_{k} \theta_k f_k(x_j, y')} \right)
\]

\[
\sum_{j=1}^{m} \left( f_{\ell}(x_j, y_j^*) - \sum_{y'} \Pr_{\hat{\theta}}(y' | x_j) f_{\ell}(x_j, y') \right)
\]

adapted from Noah A. Smith: https://homes.cs.washington.edu/~nasmith/papers/smith.tut04.pdf
The Gradient of a Log-Linear Model

\[
\sum_{j=1}^{m} \left( f_{\ell}(x_j, y_j^*) - \sum_{y'} \frac{\text{Pr}(y' | x_j)}{\hat{\theta}} f_{\ell}(x_j, y') \right)
\]

observed feature “counts”  expected feature “count”

adapted from Noah A. Smith: https://homes.cs.washington.edu/~nasmith/papers/smith.tut04.pdf
The Exponential Family

Bernoulli

Gaussian
Why “The Exponential Family”?

- The **exponential family** is a family of probability distributions over $x \in X$, parameterized by some $\theta$, of the form

$$p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta \cdot \phi(x))$$

where

- $Z(\theta)$ is the partition function
- $h(x)$ determines the support (exact zeros in the model)
- $\theta$ are the **canonical parameters**
- $\Phi(x)$ are the **sufficient statistics**
  - This is the same as a feature function! Just different terminology between statistics and NLP!
3 Why care about the Exponential Family?

- This is just one of the many ways to define the joint distribution between $x$ and $\theta$, why should you care?
- If you prove something about the exponential family, you’ve proven it about a lot of distributions at once!

...and many more
The Bernoulli is an Exponential Family Distribution

\[ p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta \cdot \phi(x)) \]

Bernoulli Distribution: Standard formulation

The Bernoulli for \( x \in \{0, 1\} \)

\[ \text{Ber}(x \mid \mu) = \mu^x (1 - \mu)^{1-x} \]

\[ = \exp \log \mu^x (1 - \mu)^{1-x} \]

\[ = \exp [\log \mu^x + \log (1 - \mu)^{1-x}] \]

\[ = \exp [x \log(\mu) + (1 - x) \log(1 - \mu)] \]
The Bernoulli is an Exponential Family Distribution

\[ p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta \cdot \phi(x)) \]

Bernoulli Distribution: Standard formulation

\[ = \exp[x \log(\mu) + (1 - x) \log(1 - \mu)] \]
\[ = \exp[\phi(x)^T \theta] \]

where \( \phi(x) = [\mathbb{I}(x = 0), \mathbb{I}(x = 1)] \) and \( \theta = [\log(\mu), \log(1 - \mu)] \).
The Bernoulli is an Exponential Family Distribution

\[ p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta \cdot \phi(x)) \]

Bernoulli Distribution: Standard formulation

\[ = \exp[x \log(\mu) + (1 - x) \log(1 - \mu)] \]

\[ = \exp[\phi(x)^T \theta] \]

where \( \phi(x) = [\mathbb{I}(x = 0), \mathbb{I}(x = 1)] \) and \( \theta = [\log(\mu), \log(1 - \mu)] \).
The Bernoulli is an Exponential Family Distribution

$$= \exp[x \log(\mu) + (1 - x) \log(1 - \mu)]$$

where $\phi(x) = [\mathbb{I}(x = 0), \mathbb{I}(x = 1)]$ and $\theta = [\log(\mu), \log(1 - \mu)]$

there is a linear dependence between the features

$$\mathbf{1}^T \phi(x) = \mathbb{I}(x = 0) + \mathbb{I}(x = 1) = 1$$

Consequently $\theta$ is not uniquely identifiable. It is common to require there is a unique $\theta$ associated with the distribution.

adapted from Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, Chapter 9: Generalized linear models and the exponential family
The Bernoulli is an Exponential Family Distribution

\[ p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta \cdot \phi(x)) \]

\textbf{Bernoulli Distribution: Standard formulation}

The Bernoulli for \( x \in \{0, 1\} \)

\[ \text{Ber}(x \mid \mu) = \mu^x (1 - \mu)^{1-x} = \mu^x (1 - \mu)(1 - \mu)^{-x} \]

\[ = (1 - \mu)(\frac{\mu}{1 - \mu})^x = (1 - \mu) \exp \log(\frac{\mu}{1 - \mu})^x \]

\[ = (1 - \mu) \exp(x \log(\frac{\mu}{1 - \mu})) \]

adapted from Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, Chapter 9: Generalized linear models and the exponential family
The Gaussian is an Exponential Family Distribution

\[ p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta \cdot \phi(x)) \]

Gaussian Distribution: Standard formulation

\[
\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{1}{2\sigma^2}(x - \mu)^2\right]
\]

\[
= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left[-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right]
\]

\[
= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\mu^2}{2\sigma^2}\right] \exp\left[-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x\right]
\]
The Gaussian is an Exponential Family Distribution

\[ p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta \cdot \phi(x)) \]

\[ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[ -\frac{\mu^2}{2\sigma^2} \right] \exp\left[ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x \right] \]

\[ Z(\theta) = \sqrt{2\pi\sigma^2} \exp\left( \frac{\mu^2}{2\sigma^2} \right) \]

\[ \theta = \left[ -\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2} \right]^T \]

\[ h(x) = 1 \]

\[ \phi(x) = [x^2, x]^T \]
Interactive Visualization
Interactive visualization

Welcome! This interactive visualization will help you understand the popular technique of log-linear modeling.

Try it out: The sliders below control the parameters ("weights") of a log-linear model. When you increase the circle weight, which filled shapes get bigger? Which ones get smaller?

One game is to try to match all 4 shapes to the gray outlines. You will need to use both sliders. A shape will turn gray if it matches well. It turns red if it is too small, blue if it is too big. Note: You may like to zoom in with your browser.

What the picture means: Your model defines a probability for each shape. You're adjusting these model probabilities by changing the weights. When the weights are 0, all 4 filled shapes have equal probability of 1/4, as shown by their equal areas.
Fin