Log-Linear Tutorial

1. Examples on random variables
2. The gradient of the log linear model
3. The exponential family
4. Interactive visualization

Examples on Random Variables

① 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; ...

$$6 \xleftarrow{\text{random variable: product}} \{2, 3\} \xrightarrow{\text{random variable: sum}} 5$$

adapted from A. Aldo Faisal, Cheng Soon Ong, and Marc Peter Deisenroth Mathematics for Machine Learning

1 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 dices.

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

2 The Gradient of a Log-Linear Model

Finding the partial derivative

$$L(\boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(y_n \mid x_n, \boldsymbol{\theta})$$

$$p(y \mid x, \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta} \cdot \mathbf{f}(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \mathbf{f}(x, y'))}$$

$$\Downarrow$$

$$\sum_{j=1}^{m} \log \Pr_{\vec{\theta}}(y_j^* \mid x_j)$$

$$\Pr_{\vec{\theta}}(y \mid x) = \frac{\exp\left(\vec{\theta} \cdot \vec{f}(x, y)\right)}{\sum_{y'} \exp\left(\vec{\theta} \cdot \vec{f}(x, y')\right)}$$

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

2 The Gradient of a Log-Linear Model

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

$$\Downarrow$$

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

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

② The Gradient of a Log-Linear Model

$$\sum_{j=1}^{m} \boxed{\log \frac{\exp\left(\vec{\theta} \cdot \vec{f}(x,y)\right)}{\sum_{y'} \exp\left(\vec{\theta} \cdot \vec{f}(x,y')\right)}}$$

$$\Downarrow$$

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

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

2 The Gradient of a Log-Linear Model

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

$$\Downarrow$$

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

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

The Gradient of a Log-Linear Model

$$\mathcal{L}\left(\vec{\theta}\right) = \left(\sum_{j=1}^{m} \boxed{\sum_{k} \theta_k f_k(x_j, y_j^*)}\right) - \sum_{j=1}^{m} \log \sum_{y'} \exp\left(\sum_{k} \theta_k f_k(x_j, y')\right)$$

Take the **derivative** with respect to θ_ℓ, we have

$$\frac{\partial \mathcal{L}}{\partial \theta_\ell} = \left(\sum_{j=1}^{m} \boxed{f_\ell(x_j, y_j^*)}\right) - \sum_{j=1}^{m} \frac{\sum_{y'} \left(\exp \sum_k \theta_k f_k(x_j, y')\right) f_\ell(x_j, y')}{\sum_{y'} \exp \sum_k \theta_k f_k(x_j, y')}$$

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

(2) The Gradient of a Log-Linear Model $\frac{\partial \mathcal{L}}{\partial \theta_\ell}$

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

$$(\ln x)' = \frac{1}{x}$$

$$\sum_{j=1}^{m} \frac{\qquad}{\sum_{y'} \exp \sum_k \theta_k f_k(x_j, y')}$$

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

② The Gradient of a Log-Linear Model $\frac{\partial \mathcal{L}}{\partial \theta_\ell}$

$$\sum_{j=1}^{m} \log \sum_{y'} \boxed{\exp\left(\sum_k \theta_k f_k(x_j, y')\right)}$$

$$(e^x)' = e^x$$

$$\sum_{j=1}^{m} \frac{\sum_{y'} (\exp \sum_k \theta_k f_k(x_j, y'))}{\sum_{y'} \exp \sum_k \theta_k f_k(x_j, y')}$$

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

② The Gradient of a Log-Linear Model $\frac{\partial \mathcal{L}}{\partial \theta_\ell}$

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

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

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

② The Gradient of a Log-Linear Model

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

$$\Downarrow$$

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

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

② The Gradient of a Log-Linear Model

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

$$\Downarrow$$

$$\sum_{j=1}^{m} \left(f_\ell(x_j, y_j^*) - \sum_{y'} \boxed{\Pr_{\vec{\theta}} (y' \mid x_j)} f_\ell(x_j, y') \right)$$

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

② The Gradient of a Log-Linear Model

$$\sum_{j=1}^{m} \left(\boxed{f_\ell(x_j, y_j^*)} - \boxed{\sum_{y'} \Pr_{\vec{\theta}}(y' \mid x_j)\, f_\ell(x_j, y')} \right)$$

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

③ Why “The Exponential Family”?

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

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

where

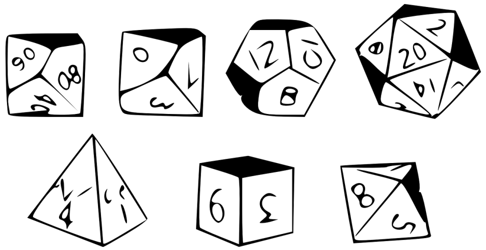
- $Z(\boldsymbol{\theta})$ is the partition function
- $h(x)$ determines the support (exact zeros in the model)
- $\boldsymbol{\theta}$ are the **canonical parameters**
- $\boldsymbol{\Phi}(x)$ are the **sufficient statistics**
  - This is the same as a feature function! Just different terminology between statistics and NLP!

adapted from Philippe Rigolett, Statistics for Applications, Lecture 21: Generalized Linear Models

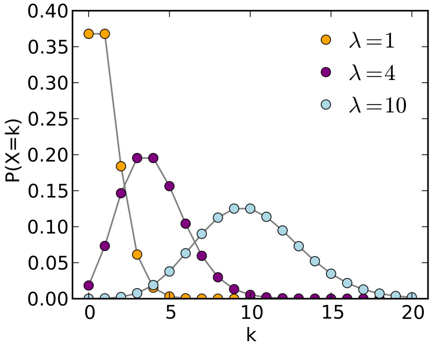
③ Why care about the Exponential Family?

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

Discrete



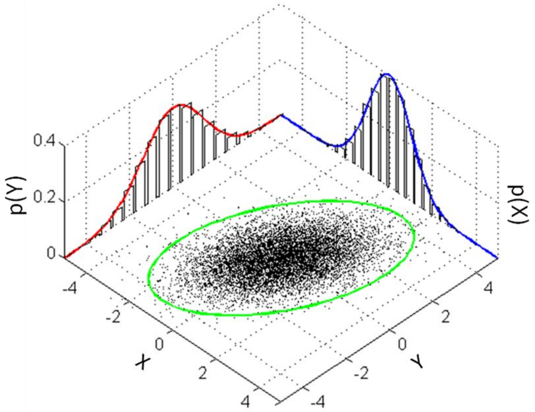
Bernoulli/Categorical



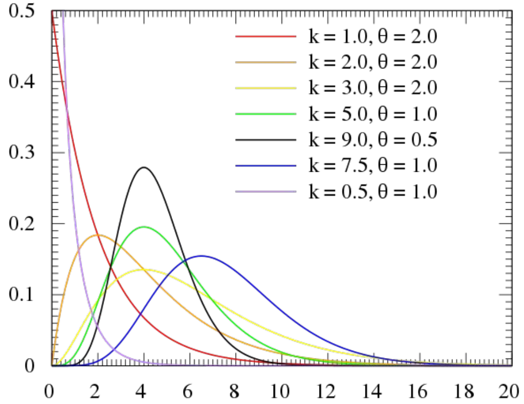
Poisson

...and many more

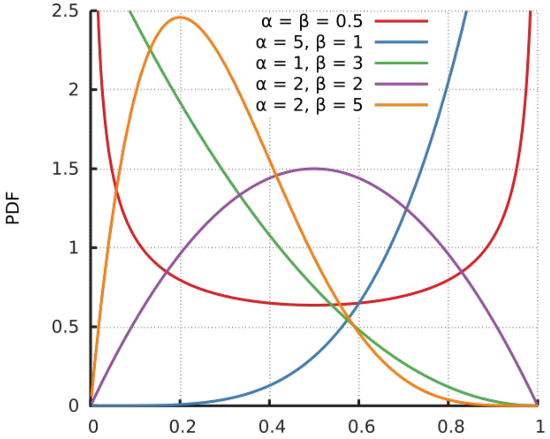
Continuous



Gaussian



Gamma



Beta

image credit: Wikipedia

③ The Bernoulli is an Exponential Family Distribution

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

Bernoulli Distribution: Standard formulation

The Bernoulli for $x \in \{0, 1\}$

$$\begin{aligned}
\mathrm{Ber}(x|\mu) &= \mu^x (1-\mu)^{1-x} \\
&= \exp \log \mu^x (1-\mu)^{1-x} \\
&= \exp[\log \mu^x + \log(1-\mu)^{1-x}] \\
&= \exp[x \log(\mu) + (1-x) \log(1-\mu)]
\end{aligned}$$

adapted from Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, Chapter 9: Generalized linear models and the exponential family

③ The Bernoulli is an Exponential Family Distribution

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

Bernoulli Distribution: Standard formulation

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

$$= \exp[\boldsymbol{\phi}(x)^T \boldsymbol{\theta}]$$

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

adapted from Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, Chapter 9: Generalized linear models and the exponential family

③ The Bernoulli is an Exponential Family Distribution

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

Bernoulli Distribution: Standard formulation

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

$$= \exp[\boldsymbol{\phi}(x)^T \boldsymbol{\theta}]$$

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

Is this representation good?

adapted from Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, Chapter 9: Generalized linear models and the exponential family

③ The Bernoulli is an Exponential Family Distribution

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

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

there is a **linear dependendence** between the features

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

Consequently **θ** is not uniquely identifiable. It is common to require there is a unique θ associated with the distribution.

adapted from Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, Chapter 9: Generalized linear models and the exponential family

③ The Bernoulli is an Exponential Family Distribution

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

Bernoulli Distribution: Standard formulation

The Bernoulli for $x \in \{0, 1\}$

$$\mathrm{Ber}(x|\mu) = \mu^x (1-\mu)^{1-x} = \mu^x (1-\mu)(1-\mu)^{-x}$$

$$= (1-\mu)\left(\frac{\mu}{1-\mu}\right)^x \quad = (1-\mu) \exp \log\left(\frac{\mu}{1-\mu}\right)^x$$

$$= (1-\mu) \exp\left(x \log\left(\frac{\mu}{1-\mu}\right)\right)$$

adapted from Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, Chapter 9: Generalized linear models and the exponential family

3 The Gaussian is an Exponential Family Distribution

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

Gaussian Distribution: Standard formulation

$$\begin{aligned}
\mathcal{N}(x|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2] \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp[-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x \boxed{- \frac{1}{2\sigma^2}\mu^2}]
\end{aligned}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \boxed{\exp[-\frac{\mu^2}{2\sigma^2}]} \exp[-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x]$$

3 The Gaussian is an Exponential Family Distribution

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

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

$$Z(\theta) = \sqrt{2\pi\sigma^2} \exp(\frac{\mu^2}{2\sigma^2})$$

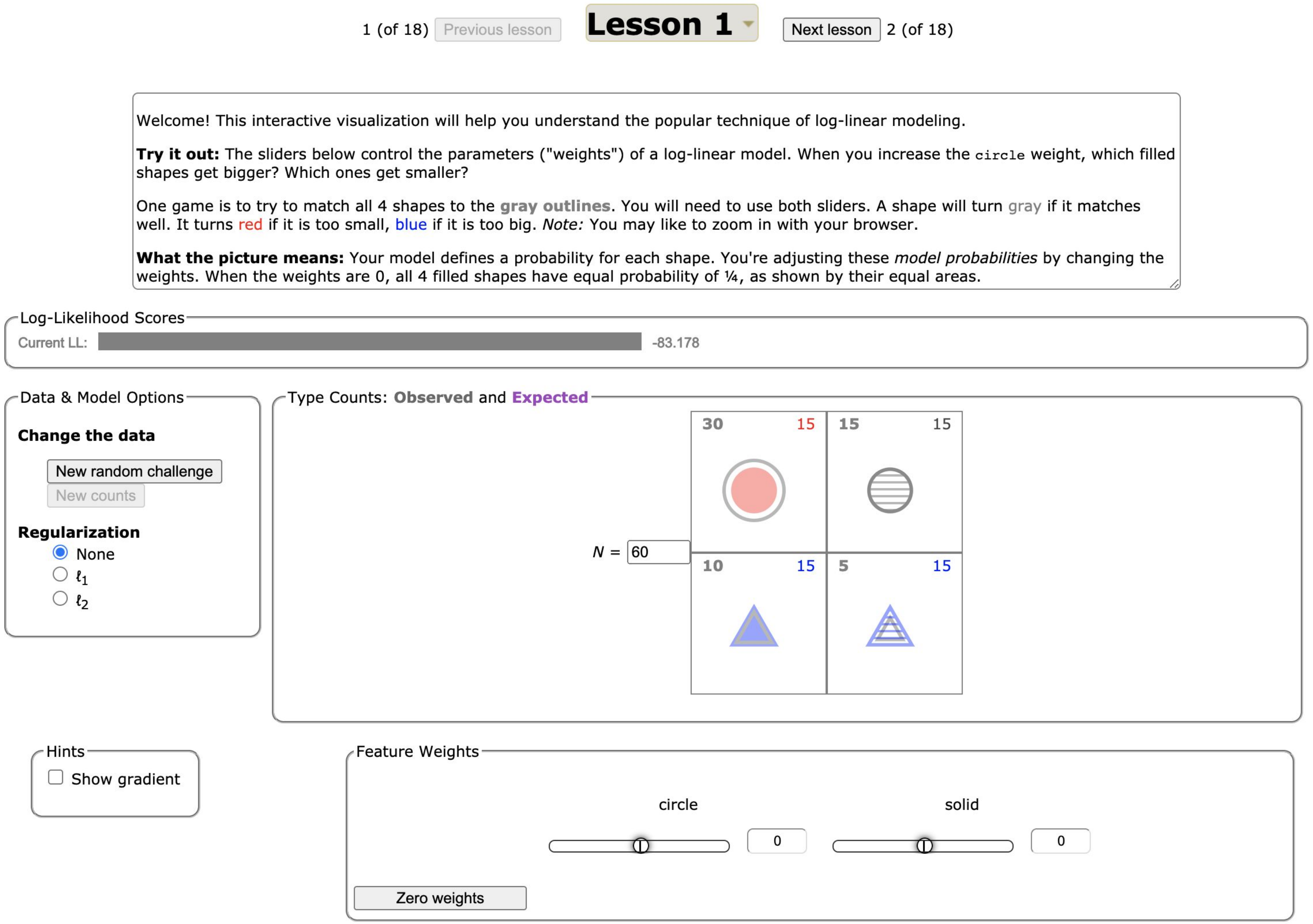
$$\theta = [-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}]^T$$

$$h(x) = 1$$

$$\phi(x) = [x^2, x]^T$$

Interactive Visualization

④ Interactive visualization



ETH zürich

Fin