DSC291: Advanced Statistical Natural Language Processing

Machine Learning Basics

Zhiting Hu

Lecture 2, March 31, 2022

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

- Probability
  - Bayes' rule
  - Exponential family
  - KL divergence, cross entropy
- Functional derivatives (optional)
- Practice: MLE vs Maximum entropy

Probability

Why Probability?

- The world is a very uncertain place
  - "What will the weather be like today?"
  - "Will I like this movie?"
- We often can't prove something is true, but we can still ask how likely different outcomes are or ask for the most likely explanations
- Predictions need to have associated confidence
  - Confidence -> probability
- Not all machine learning models are probabilistic
  - … but most of them have probabilistic interpretations

[CS60020, Bhattacharya; CSC2515, Wang]

Notations

- A random variable $\boldsymbol{x}$ represents outcomes or states of the world.
  - We write $p(\boldsymbol{x}_0)$ to mean Probability($\boldsymbol{x} = \boldsymbol{x}_0$)
- Sample space: the space of all possible outcomes (may be discrete, continuous, or mixed)
- $p(\boldsymbol{x})$ is the probability mass (density) function
  - Assigns a number to each point in sample space
  - Non-negative, sums (integrates) to 1
  - Intuitively: how often does $\boldsymbol{x}$ occur, how much do we believe in $\boldsymbol{x}$.

Notations

- Joint distribution $p(\boldsymbol{x}, \boldsymbol{y})$
- Conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$
  - $p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})}$
- Expectation:

$$\mathbb{E}[f(\boldsymbol{x})] = \sum_{\boldsymbol{x}} f(\boldsymbol{x})\, p(\boldsymbol{x})$$

or

$$\mathbb{E}[f(\boldsymbol{x})] = \int_{\boldsymbol{x}} f(\boldsymbol{x}) p(\boldsymbol{x}) dx$$

Rules of Probability

- Sum rule

$$p(x) = \sum_y p(x, y)$$ (Marginalize out y)

$$p(x_1) = \sum_{x_2} \sum_{x_3} \cdots \sum_{x_N} p(x_1, x_2, \ldots, x_N)$$

- Product/chain rule

$$p(x, y) = p(y \mid x) p(x)$$

$$p(x_1, \ldots, x_N) = p(x_1) p(x_2 \mid x_1) \ldots p(x_N \mid x_1, \ldots, x_{N-1})$$

Bayes' Rule

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

- This gives us a way of "reversing" conditional probabilities
- We call $p(\boldsymbol{y})$ the "prior", and $p(\boldsymbol{y}|\boldsymbol{x})$ the "posterior"
- Ex: Bayes' Rule in machine learning:
  - $\mathcal{D}$: data (evidence)
  - $\boldsymbol{\theta}$: unknown quantities, such as model parameters, predictions

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

Posterior belief on the unknown quantities you see data $\mathcal{D}$

Likelihood: How likely is the observed data under the particular unknown quantities $\boldsymbol{\theta}$

Prior belief on the unknown quantities **Before** you see data $\mathcal{D}$

Independence

- Two random variables are said to be **independent** iff their joint distribution factors

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$$

- Two random variables are **conditionally independent** given a third if they are independent after conditioning on the third

$$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{y}|\boldsymbol{z})$$

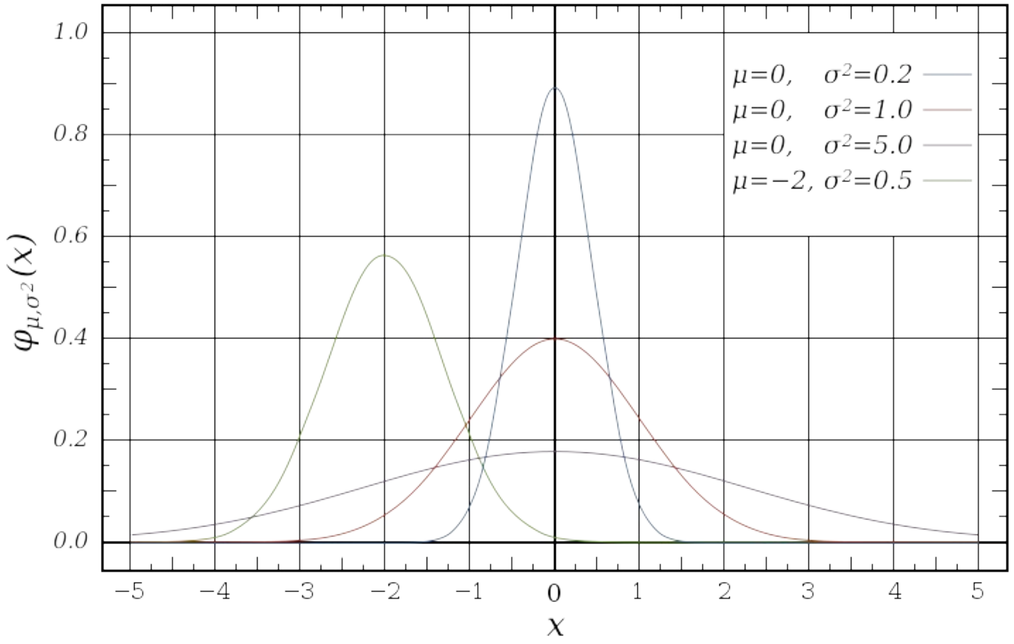
Some common distributions - Gaussian distribution

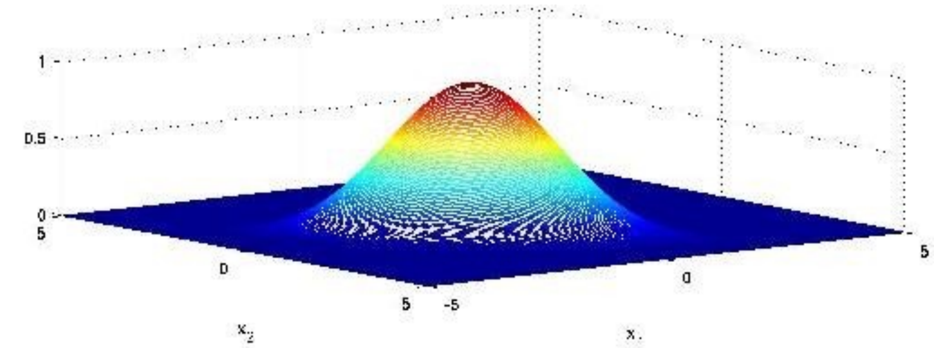
- Gaussian distribution

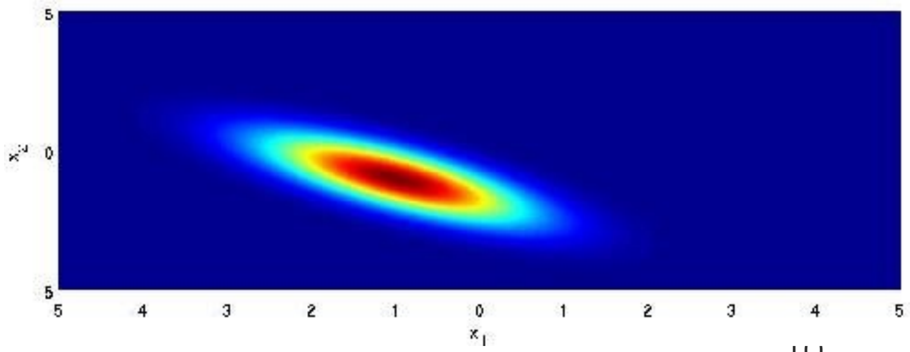
$$P(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$

(Multivariate)

$$P(x \mid \mu, \Sigma) = \left|2\pi \Sigma\right|^{-1/2} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right\}$$







[CSC2515, Wang]

Some common distributions - Multinomial distribution

- Multinomial distribution
  - Discrete random variable $\boldsymbol{x}$ that takes one of M values $\{1, \ldots, M\}$
  - $p(\boldsymbol{x} = i) \ = \ \pi_i, \qquad \sum_i \pi_i = 1$
  - Out of n independent trials, let k_i be the number of times $\boldsymbol{x} = i$ was observed
  - The probability of observing a vector of occurrences $\boldsymbol{k} = [k_1, \ldots, k_M]$ is given by the *multinomial distribution* parametrized by $\boldsymbol{\pi}$

$$p(\mathbf{k}|\boldsymbol{\pi}, n) \ = \ p(k_1, \ldots, k_m|\pi_1, \ldots, \pi_m, n) \ = \ \frac{n!}{k_1!k_2!\ldots k_m!}\prod_{i=1}\pi_i^{k_i}$$

  - E.g., describing a text document by the frequency of occurrence of every distinct word
  - For $n = 1$, a.k.a. categorical distribution
    - $p(\boldsymbol{x} = i \mid \boldsymbol{\pi}) \ = \ \pi_i$
    - In $\boldsymbol{k} = [k_1, \ldots, k_M]$: $k_i = 1$, and $k_j = 0$ for all $j \neq i \quad \rightarrow \quad a.k.a.,$ one-hot representation of i

Exponential family

- A distribution

$$p_{\theta}(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\} / Z(\boldsymbol{\theta})$$

  is an exponential family distribution

  - $\boldsymbol{\theta} \in R^d$: natural (canonical) parameter
  - $T(\boldsymbol{x}) \in R^d$: sufficient statistics, features of data $\boldsymbol{x}$
  - $Z(\boldsymbol{\theta}) = \sum_{x,y} h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\}$: normalization factor

- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma,...

Example: Multivariate Gaussian Distribution

- For a continuous vector random variable $\boldsymbol{x} \in R^k$

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

Moment parameter

$$= \frac{1}{(2\pi)^{k/2}} \exp\left\{-\tfrac{1}{2}\mathrm{tr}\left(\Sigma^{-1}xx^T\right) + \mu^T\Sigma^{-1}x - \tfrac{1}{2}\mu^T\Sigma^{-1}\mu - \log|\Sigma|\right\}$$

- Exponential family representation

$$\boldsymbol{\theta} = \left[\Sigma^{-1}\mu; -\frac{1}{2}\mathrm{vec}\left(\Sigma^{-1}\right)\right] = \left[\boldsymbol{\theta}_1, \mathrm{vec}\left(\boldsymbol{\theta}_2\right)\right], \; \boldsymbol{\theta}_1 = \Sigma^{-1}\mu \text{ and } \boldsymbol{\theta}_2^{-} = -\frac{1}{2}\Sigma^{-1}$$

$$T(\boldsymbol{x}) = \left[\boldsymbol{x}; \mathrm{vec}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right]$$

$$A(\boldsymbol{\theta}) = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \log|\Sigma| = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\theta}_2\boldsymbol{\theta}_1\boldsymbol{\theta}_1^T\right) - \frac{1}{2}\log\left(-2\boldsymbol{\theta}_2\right)$$

$$h(x) = (2\pi)^{-k/2}$$

Entropy

- Shannon entropy $H(p) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x})$
  - The average level of "information", "surprise", or "uncertainty" inherent to the variable $\boldsymbol{x}$'s possible outcomes

KL Divergence

- Kullback-Leibler (KL) divergence: measures the closeness of two distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$

$$\mathrm{KL}(q(\boldsymbol{x}) \,||\, p(\boldsymbol{x})) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$$

  - a.k.a. Relative entropy
  - KL >= 0 (Jensen's inequality)
  - Intuitively:
    - If q is high and p is high, then we are happy (i.e. low KL divergence)
    - If q is high and p is low then we pay a price (i.e. high KL divergence).
    - If q is low then we don't care (i.e. also low KL divergence, regardless of p)
  - not a true "distance":
    - not commutative (symmetric) $\mathrm{KL}(\mathrm{p}||\mathrm{q}) \,!= \mathrm{KL}(\mathrm{q}||\mathrm{p})$
    - doesn't satisfy triangle inequality

KL Divergence

- Kullback-Leibler (KL) divergence: measures the closeness of two distributions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$

$$\mathrm{KL}(q(\boldsymbol{x}) \,||\, p(\boldsymbol{x})) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$$

  - a.k.a. Relative entropy

- Maximum likelihood estimation (MLE) is minimizing the KL divergence between the empirical data distribution and the model distribution

$$\mathrm{KL}(\tilde{p}(\boldsymbol{x}) \,||\, p_\theta(\boldsymbol{x})) = \underbrace{-\mathbb{E}_{\tilde{p}(\boldsymbol{x})}[\, \log p_\theta(\boldsymbol{x}) \,]}_{\text{Cross entropy}} + H(\tilde{p}(\boldsymbol{x}))$$

Key Takeaways

- Probability $p(\boldsymbol{x})$
- Bayes' rule $$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y})}{p(\boldsymbol{x})}$$
  - prior, posterior
- Exponential family:
  - Gaussian, multinomial, categorical, …

$$\mathrm{KL}(q(\boldsymbol{x}) \,||\, p(\boldsymbol{x})) = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}$$

- KL Divergence
  - relation to Cross-entropy

Functional Derivatives (optional)

Functional derivative

- $\nabla_q - \mathbb{H}(q) = \log q + 1$

- Functional $F(y)$: an operator that takes a function $y(x)$ and returns an output value F
- Functional derivative (aka, variational derivative): relates a change in a Functional $F(y)$ to a change in the function y

Functional derivative

- Recall the conventional derivative $\frac{dy}{dx}$
  - Taylor expansion

$$y(x+\epsilon) = y(x) + \frac{\mathrm{d}y}{\mathrm{d}x}\epsilon + O(\epsilon^2)$$

- Functional derivative
  - How much a functional $F[y]$ changes when we make a small change $\varepsilon\eta(x)$ to the function $y(x)$

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x)\,\mathrm{d}x + O(\epsilon^2)$$

  - A function $y(x)$ that maximizes (or minimizes) a functional $F[y]$ must satisfy

$$\frac{\delta F}{\delta y(x)} = 0 \text{ for all } x$$

Functional derivative

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x) \, \mathrm{d}x + O(\epsilon^2)$$

- Consider a functional that is defined by an integral over a function $G(y, x)$

$$F[y] = \int G(y, x) dx$$

- Consider variations in the function $y(x)$,

$$F[y + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\partial G}{\partial y} \eta(x) dx + O(\epsilon^2)$$

Functional derivative

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)} \eta(x)\,\mathrm{d}x + O(\epsilon^2)$$

- Consider a functional that is defined by an integral over a function $G(y, x)$

$$F[y] = \int G(y, x) dx$$

  - Ex.1, $-\mathbb{H}(q) = \int q(x) \log q(x)\, dx$
    - $G = q(x) \log q(x)$
- Consider variations in the function $y(x)$,

$$F[y + \epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\partial G}{\partial y} \eta(x) dx + O(\epsilon^2)$$

Practice: Maximum likelihood vs Maximum Entropy

Supervised Maximum Likelihood

- Model to be learned $p_\theta(\boldsymbol{x})$
- Observe full data $\mathcal{D} = \{ \boldsymbol{x}^* \}$
  - i.i.d: independent, identically distributed

- Maximum Likelihood Estimation (MLE)
  - The most classical learning algorithm

$$\min_{\theta} - \mathbb{E}_{x^* \sim \mathcal{D}} \left[\log p_\theta(\boldsymbol{x}^*) \right]$$

- MLE is closely connected to the Maximum Entropy (MaxEnt) principle

Recap: Exponential Family

- A distribution

$$p_\theta(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\} / Z(\boldsymbol{\theta})$$

  is an exponential family distribution

  - $\boldsymbol{\theta} \in R^d$: natural (canonical) parameter
  - $T(\boldsymbol{x}) \in R^d$: sufficient statistics, features of data $\boldsymbol{x}$
  - $Z(\boldsymbol{\theta}) = \sum_{x,y} h(\boldsymbol{x}) \exp\{\boldsymbol{\theta} \cdot T(\boldsymbol{x})\}$: normalization factor

- Examples: Bernoulli, multinomial, Gaussian, Poisson, gamma,...

Maximum Likelihood for Exponential Family

$m(\boldsymbol{x})$: the number of times $\boldsymbol{x}$ is observed in *D*

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \log p(\boldsymbol{x} \mid \boldsymbol{\theta})$$

$$= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \left(\sum_i \theta_i T_i(\boldsymbol{x}) - \log Z(\boldsymbol{\theta}) \right)$$

$$= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \sum_i \theta_i T_i(\boldsymbol{x}) - N \log Z(\boldsymbol{\theta})$$

- Take gradient and set to 0

$$\frac{\partial}{\partial \theta_i} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) = \sum_{\boldsymbol{x}} m(\boldsymbol{x}) T_i(\boldsymbol{x}) - N \frac{\partial}{\partial \theta_i} \log Z(\boldsymbol{\theta})$$

$$= \sum_{\boldsymbol{x}} m(\boldsymbol{x}) T_i(\boldsymbol{x}) - N \sum_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{\theta}) T_i(\boldsymbol{x})$$

$$\Rightarrow \sum_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{\theta}) T_i(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \frac{m(\boldsymbol{x})}{N} T_i(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \tilde{p}(\boldsymbol{x} \mid \boldsymbol{\theta}) T_i(\boldsymbol{x})$$

At MLE, the expectations of the sufficient statistics under the model must match empirical feature average

Maximum Entropy (MaxEnt)

- Given $\mathcal{D}$, to estimate $p(\boldsymbol{x})$
- We can approach the problem from an entirely different point of view. Begin with some fixed feature expectations:

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \frac{m(\boldsymbol{x})}{N} T_i(\boldsymbol{x}) := \alpha_i$$

- There may exist many distributions which satisfy them. Which one should we select?
  - MaxEnt principle: the most uncertain or flexible one, i.e., the one with maximum entropy
- This yields a new optimization problem:
  - This is a variational definition of a distribution!

$$\max_{p} \; \mathrm{H}(p(\boldsymbol{x})) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x})$$

$$\text{s.t.} \quad \sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \alpha_i$$

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} \; L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_i \theta_i \left(\sum_x p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} \; L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_i \theta_i \left(\sum_x p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

$$\frac{\partial L}{\partial p(\boldsymbol{x})} = 1 + \log p(\boldsymbol{x}) - \sum_i \theta_i T_i(\boldsymbol{x}) - \mu$$

$$p^*(\boldsymbol{x}) = e^{\mu - 1} \exp \left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\}$$

$$Z(\boldsymbol{\theta}) = e^{\mu - 1} = \sum_{\boldsymbol{x}} \exp \left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\} \quad \left(\text{since } \sum_{\boldsymbol{x}} p^*(\boldsymbol{x}) = 1 \right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_i \theta_i \left(\sum_x p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

$$\frac{\partial L}{\partial p(\boldsymbol{x})} = 1 + \log p(\boldsymbol{x}) - \sum_i \theta_i T_i(\boldsymbol{x}) - \mu$$

$$p^*(\boldsymbol{x}) = e^{\mu - 1} \exp\left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\}$$

$$Z(\boldsymbol{\theta}) = e^{\mu - 1} = \sum_{\boldsymbol{x}} \exp\left\{ \sum_i \theta_i f_i(\boldsymbol{x}) \right\} \quad \left(\text{since} \sum_{\boldsymbol{x}} p^*(\boldsymbol{x}) = 1 \right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

- So feature constraints + MaxEnt ⇒ exponential family.
- Problem is strictly convex w.r.t. $p(\boldsymbol{x})$, so solution is unique.

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$\max_{\theta,\mu} \min_{p(x)} L = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) - \sum_i \theta_i \left(\sum_x p(\boldsymbol{x}) T_i(\boldsymbol{x}) - \alpha_i \right) - \mu \left(\sum_{\boldsymbol{x}} p(\boldsymbol{x}) - 1 \right)$$

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

plug $p(x|\boldsymbol{\theta})$ back into L, and since $\sum_{\boldsymbol{x}} \frac{m(\boldsymbol{x})}{N} T_i(\boldsymbol{x}) \coloneqq \alpha_i$:

$$\max_{\theta} L(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} m(\boldsymbol{x}) \sum_i \theta_i T_i(\boldsymbol{x}) - N \log Z(\boldsymbol{\theta})$$

- Recovers precisely the MLE problem of exponential family

(Homework)

- So feature constraints + MaxEnt $\Rightarrow$ exponential family.
- Problem is strictly convex w.r.t. $p(\boldsymbol{x})$, so solution is unique.

Constraints from Data

- We have seen a case of **convex duality**:
  - In one case, we assume exponential family and show that Maximum Likelihood implies model expectations must match empirical expectations.
  - In the other case, we assume model expectations must match empirical feature counts and show that MaxEnt implies exponential family distribution.

A more general MaxEnt problem

$$\min_p \quad \mathrm{KL}(p(\boldsymbol{x}) \| h(\boldsymbol{x}))$$

$$\stackrel{\text{def}}{=} \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{h(\boldsymbol{x})} = -\mathrm{H}(p) - \sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log h(\boldsymbol{x})$$

$$\text{s.t.} \quad \sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \alpha_i$$

$$\sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$$

$$\Rightarrow \quad p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\boldsymbol{x}) \exp \left\{ \sum_i \theta_i T_i(\boldsymbol{x}) \right\}$$

Summary

- Maximum entropy is dual to maximum likelihood of exponential family distributions
- This provides an alternative view of the problem of fitting a model into data:
    - The data instances in the training set are treated as constraints, and the learning problem is treated as a constrained optimization problem.
    - We'll revisit this optimization-theoretic view of learning repeatedly in the future!

$$
\begin{aligned}
\max_{p} \quad & \mathrm{H}(p(\boldsymbol{x})) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) \\
\text{s.t.} \quad & \sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \alpha_i \\
& \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1
\end{aligned}
$$

Key Takeaways

- Probability
  - Bayes’ rule
  - Exponential family
  - KL divergence
- Functional derivative (optional, but very useful)
- Convex duality between MLE and MaxEnt (optional)

Questions?