DSC250: Advanced Data Mining

Machine Learning Basics

Zhiting Hu

Lecture 5, Jan 21, 2025

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Outline

- Probability
  - Bayes' rule
  - Exponential family
  - Probabilistic graphical models
  - Entropy, KL divergence, cross entropy
- Functional derivatives
- Practice: MLE vs Maximum entropy

High dimensional distributions

Probabilistic graphical models (PGMs) are about representing probability distributions over random variables

$$p(X) \equiv p(X_1, \ldots, X_n)$$

Assume $x_i \in \{0, 1\}^n$

Naively, since there are 2^n possible assignments to $X_1, \ldots, X_n$, can represent this distribution completely using $2^n - 1$ numbers, but quickly becomes intractable for large n

PGMs are methods to represent these distributions more compactly, by exploiting *conditional independence*

Recap: Bayesian networks (directed PGMs)

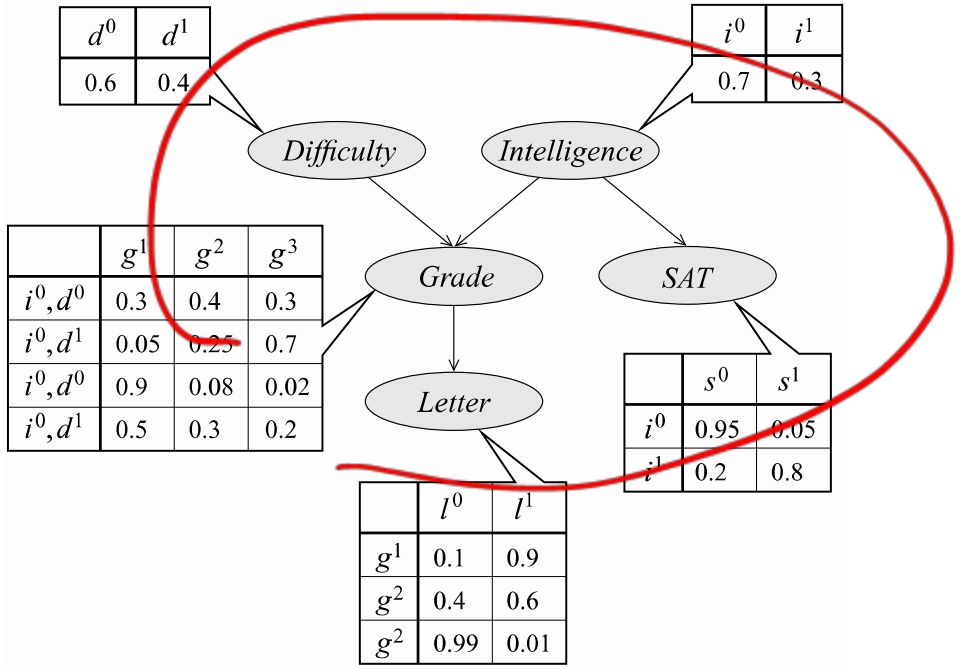
- A **Bayesian network** is specified by a directed *acyclic* graph $G = (V, E)$ with:
  1. One node $i \in V$ for each random variable X_i
  2. One conditional probability distribution (CPD) per node, $p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values
- Corresponds 1-1 with a particular factorization of the joint distribution:

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

- Powerful framework for designing *algorithms* to perform probability computations

Recap: Example

- Consider the following Bayesian network:



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^0, d^0	0.9	0.08	0.02
i^0, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

- What is its joint distribution?

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

$$p(d, i, g, s, l) = p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)$$

Entropy

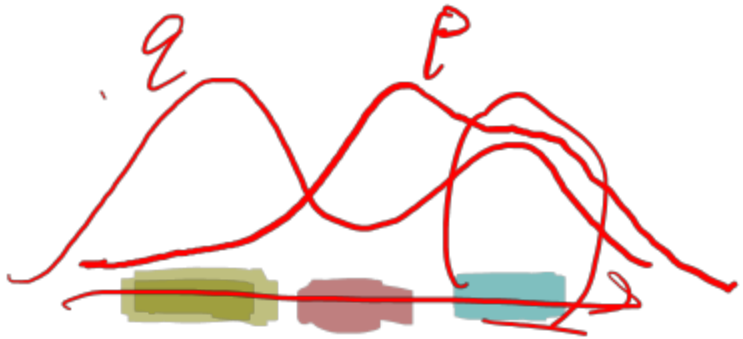
Entropy

- Shannon entropy $H(p) = -\sum_{x} p(\boldsymbol{x}) \log p(\boldsymbol{x})$
  - 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 $q(\boldsymbol{x})$

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



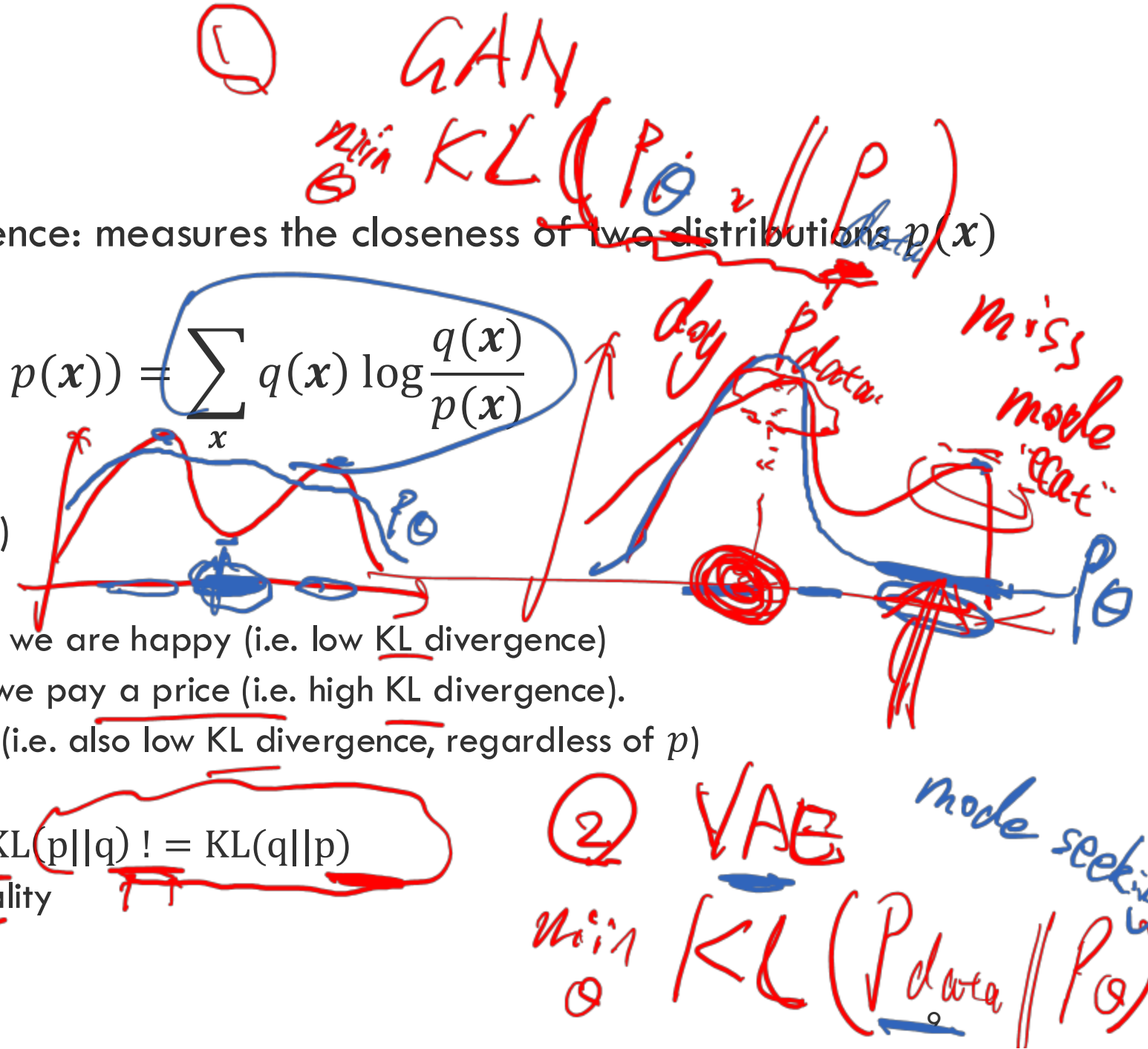
  - a.k.a. Relative entropy
  - KL >= 0 (Jensen's inequality)
  - **Questions:**
    - If q is high and p is high in a region, then KL divergence is _____ in this region.
    - If q is high and p is low in a region, then KL divergence is _____ in this region.
    - If q is low in a region, then KL divergence is _____ in this region.

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)

$$\min_{\theta} - \mathbb{E}_{x \sim \tilde{p}(x)} \left[\log p_\theta(\boldsymbol{x}) \right]$$

- **Question:** Show that MLE is minimizing the KL divergence between the empirical data distribution and the model distribution

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)

$$\min_{\theta} - \mathbb{E}_{x \sim \tilde{p}(x)} \left[\log p_\theta(\boldsymbol{x}) \right]$$

- **Question:** Show that MLE is minimizing the KL divergence between the empirical data distribution and the model distribution

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

Cross entropy

Key Takeaways

- Probability $p(\boldsymbol{x})$
- Bayes' rule
  - prior, posterior

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

- Exponential family:
  - Gaussian, multinomial, categorical, …
- Probabilistic graphical models: Bayesian networks
- KL Divergence
  - relation to Cross-entropy

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

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 $\frac{dy}{dx}$
  - How much a function $y(x)$ changes when we make a small change Δx to the input x
  - 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 $G(y,x)$

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

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

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

$$\frac{\partial q \cdot \log q}{\partial q} = \log q + q \cdot \frac{1}{q} = \log q + 1 \qquad \frac{\partial H(q)}{\partial q} = \frac{\partial G}{\partial q}$$

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 (Optional)

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!

$$\max_{p} \ \mathrm{H}(p(\boldsymbol{x}))$$

$$\text{s.t.} \begin{cases} \sum_{\boldsymbol{x}} p(\boldsymbol{x}) T_i(\boldsymbol{x}) = \alpha_i \\ \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1 \end{cases}$$

stdm is defined as a solution to a opt problem

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 $\Rightarrow$ 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:

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

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

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

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

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

- Recovers precisely the MLE problem of exponential family

(Homework)

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!

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

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

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

Key Takeaways

- Probability
  - Bayes' rule
  - Exponential family
  - Probabilistic graphical models: Bayesian networks
  - KL divergence
- Functional derivative
- Convex duality between MLE and MaxEnt (optional)

Questions?