Maximum entropy models

Daniel Hsu

COMS 6998-7 Spring 2025

Statistical modeling and maximum entropy

Statistical modeling

(Berger, Della Pietra, Della Pietra, 1996)

Statistical modeling addresses the problem of constructing a stochastic model to predict the behavior of a random process. In constructing this model, we typically have at our disposal a sample of output from the process. Given this sample, which represents an incomplete state of knowledge about the process, the modeling problem is to parlay this knowledge into a representation of the process. We can then use this representation to make predictions about the future behavior about the process.

Statistical modeling for machine translation

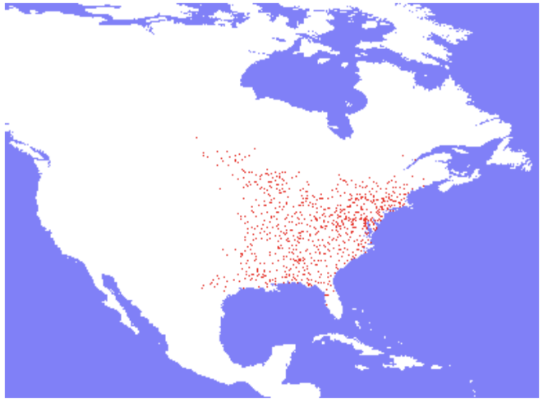
- What's the correct French translation of the English word "in"?
  - If you don't know French, all French words might seem equally plausible
- **Statistical machine translation**: Use data to find the translation
- **Data**: you see translations produced by an expert
- **Observation 1**: it is always translated to a word from the set
  { dans, en, à, au cours de, pendant }
- **Observation 2**: 30% of the times, the translation is from the set
  { dans, en }
- **Observation 3**: (something about context around English word "in")
- …

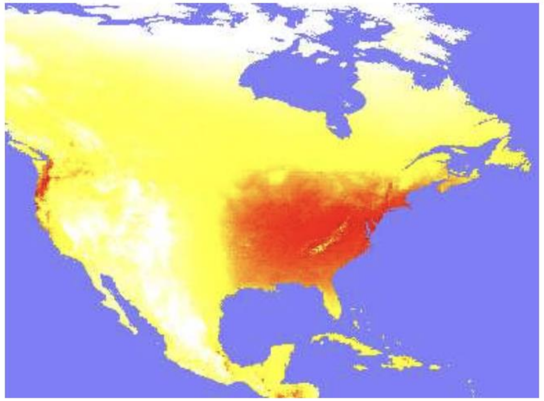
Statistical modeling for species distributions

(Phillips, Dudík, Schapire, 2004)

Where in North America do we find the Yellow-throated Vireo (YV)?

- *A priori*: all locations in North America seem equally likely to me
- **Data**: locations of YV sightings in North America
- Also have **environmental measurements** for all North American locations (e.g., annual rainfall, average daily temperature, elevation)
- **Goal**: Construct distribution over North American locations that agrees with the environmental measurements of locations where YV was sighted





General problem setup

- Finite domain $\mathcal{X}$ (e.g., all locations in North America)
  - Let q_0 be the "default model" you would've picked before seeing any data (e.g., $q_0 =$ uniform distribution on $\mathcal{X}$), a.k.a. "base measure"
- Measure some "features" of the information source
  - Get average (i.e., expected) values of n "feature functions"
$$T_i : \mathcal{X} \to \mathbb{R}$$
  - Example:
$$T_1(x) = \text{annual rainfall (in inches) at } x$$
$$T_2(x) = \mathbb{I}\{x \text{ is in the forest}\}$$
  - Let b_i be the average value of T_i in the information source
  - Default model q_0 may not be consistent with these measurements!
  - So what model should you choose instead?

Maximum entropy (maxent) principle

Maxent principle: Choose model as close to default model as possible while being consistent with measurements

$$\min_{p\in\Delta} \mathrm{RE}(p, q_0)$$
$$\text{s.t. } p[T_i] = b_i \;\; \forall i = 1, \dots, n$$

New notation:

$$p[f] \coloneqq \sum_{x\in\mathcal{X}} p(x) f(x)$$

- Recall: $\mathrm{RE}(p, q) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = p\left[\log \frac{p}{q}\right]$
- If q_0 is uniform, then $\mathrm{RE}(p, q_0) = -H(p) + \log|\mathcal{X}|$ (hence "maxent")
- Objective function is strictly convex, and constraints are linear!

Form of maxent solutions

Theorem: Whenever the maxent problem is feasible (and excluding a measure-zero set of $(b_1, \dots, b_n)$), the solution has the form

$$p_\lambda(x) = \frac{1}{Z(\lambda)} \exp\left(\sum_{i=1}^{n} \lambda_i T_i(x)\right) q_0(x)$$

for some "parameter vector" $\lambda = (\lambda_1, \dots, \lambda_n)$, where

$$Z(\lambda) = \sum_{x \in \mathcal{X}} \exp\left(\sum_{i=1}^{n} \lambda_i T_i(x)\right) q_0(x)$$

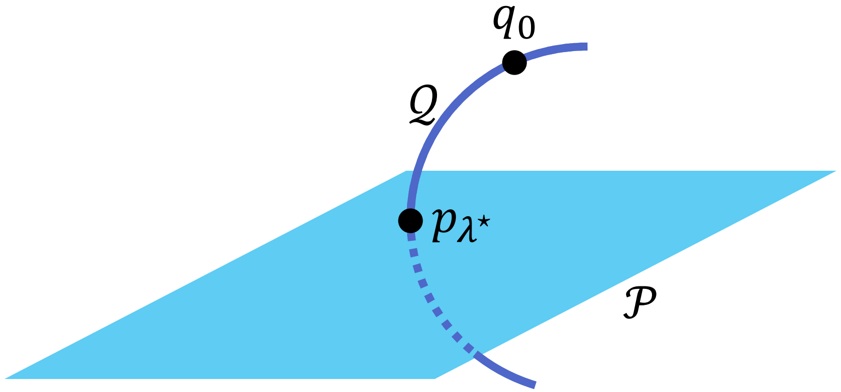
- Distributions of this form are called **Gibbs** or **Boltzmann distributions**
- Also related to **exponential families** (where q_0 need not be probability dist.)

Gibbs distributions

- The Gibbs distributions (corresponding to $T_1, \ldots, T_n$ and q_0) form a parametric family of distributions $\{p_\lambda : \lambda \in \mathbb{R}^n\}$
- Each p_λ is an "exponential tilting" of the base measure q_0
  - Suppose $T_2(x) = \mathbb{I}\{x \text{ is in the forest}\}$ and $\lambda_2 = -2.1$
  - Then a location in the forest is $\exp(-2.1) \approx 0.12$ as likely (according to p_λ) as a location not in the forest (all else being equal):

$$\frac{p_\lambda(x)}{p_\lambda(y)} = \frac{\exp(\lambda_1 T_1(x) + \lambda_2 T_2(x) + \cdots)}{\exp(\lambda_1 T_1(y) + \lambda_2 T_2(y) + \cdots)}$$

Geometric interpretation



- Notation:
  - $T(x) = (T_1(x), \ldots, T_n(x))$
  - $(\lambda \cdot T)(x) = \lambda_1 T_1(x) + \cdots + \lambda_n T_n(x)$
  - $b = (b_1, \ldots, b_n)$
- Feasible set: $\mathcal{P} = \{p \in \Delta : p[T] = b\}$, an affine set
- Maxent problem: Find $p \in \mathcal{P}$ that minimizes $\mathrm{RE}(p, q_0)$
  - Like "projection" of q_0 onto $\mathcal{P}$, except notion of "distance" is relative entropy
- Gibbs distributions (based on T, q_0): $\mathcal{Q} = \{p_\lambda : \lambda \in \mathbb{R}^n\}$
- It turns out whenever $\mathcal{P} \neq \emptyset$, then **maxent solution** is the unique distribution in both $\mathcal{P}$ and (the closure of) $\mathcal{Q}$

Deriving the form of maxent solutions

Method of Lagrange multipliers

- Maxent: Find $p \in \mathcal{P} = \{p \in \Delta : p[T] = b\}$ that minimizes $\mathrm{RE}(p, q_0)$
- To each constraint $p[T_i] = b_i$, associate a **Lagrange multiplier** λ_i
- **Lagrangian function**: for $\lambda = (\lambda_1, \dots, \lambda_n)$

$$\mathcal{L}(p, \lambda) = \mathrm{RE}(p, q_0) - \overbrace{\sum_{i=1}^{n} \lambda_i (p[T_i] - b_i)}^{\text{Affine in } \lambda}$$

$$= \underbrace{\mathrm{RE}(p, q_0) - p[\lambda \cdot T]}_{\text{Convex in } p} + \lambda \cdot b$$

- Maxent problem is

$$\min_{p \in \Delta} \sup_{\lambda \in \mathbb{R}^n} \mathcal{L}(p, \lambda)$$

Convex duality

Maxent problem satisfies conditions for a **minmax** theorem:

$$\min_{p\in\Delta}\sup_{\lambda\in\mathbb{R}^n}\mathcal{L}(p,\lambda) = \sup_{\lambda\in\mathbb{R}^n}\min_{p\in\Delta}\mathcal{L}(p,\lambda)$$

Dual objective function

$$\lambda \mapsto \min_{p\in\Delta}\mathcal{L}(p,\lambda)$$

Question: For fixed λ, what $p\in\Delta$ minimizes $\mathcal{L}(p,\lambda)$?

Donsker-Varadhan inequality: for any $f:\mathcal{X}\to\mathbb{R}$ and all $p,q\in\Delta$

$$\mathrm{RE}(p,q) \geq p[f] - \log q[\exp(f)]$$

- So $\mathcal{L}(p,\lambda) \geq -\log q_0[\exp(\lambda\cdot T)] + \lambda\cdot b$
- Furthermore, $\mathcal{L}(p_\lambda,\lambda) = -\log q_0[\exp(\lambda\cdot T)] + \lambda\cdot b$ Dual objective function

If $\lambda^\star$ maximizes dual objective, then $p_{\lambda^\star}$ is maxent solution

Connection to maximum likelihood estimation

- Suppose b is empirical average of T on data set $x^1, \dots, x^m \in \mathcal{X}$

$$b = \frac{1}{m}\sum_{j=1}^{m} T(x^j)$$

- Consider family of Gibbs distributions $\mathcal{Q}$; how to estimate parameter λ?
- Log-likelihood of p_λ (treating data set as i.i.d. sample) is

$$\log \prod_{j=1}^{m} \frac{p_\lambda(x^j)}{q_0(x^j)} = \cdots = m(-\ln q_0[\exp(\lambda \cdot T)] + \lambda \cdot b)$$

Dual objective function!

- Maximum likelihood estimation for Gibbs distributions = maximum entropy

Recap (so far)

The following are equivalent (for essentially all b):

- Distribution p that minimizes $\mathrm{RE}(p, q_0)$ subject to $p[T] = b$
- Gibbs distribution

  $$p_\lambda(x) = \frac{1}{Z(\lambda)} \exp\big((\lambda \cdot T)(x)\big)\, q_0(x)$$

  satisfying $p_\lambda[T] = b$
- Maximum likelihood Gibbs distribution p_λ (when $b = \frac{1}{m} \sum_{j=1}^{m} T\big(x^j\big)$)

Log partition function

Log partition function

- Normalization quantity used to ensure p_λ is a probability distribution

$$Z(\lambda) = \sum_{x \in \mathcal{X}} \exp\big((\lambda \cdot T)(x)\big)\, q_0(x)$$

is also called **partition function**

  - Can also write as $Z(\lambda) = q_0[\exp(\lambda \cdot T)]$
  - Can also interpret as **moment generating function** for $T(X)$ where $X \sim q_0$

- Logarithm of partition function is called _______________

$$G(\lambda) = \log Z(\lambda) = \log q_0[\exp(\lambda \cdot T)]$$

- Can write

$$p_\lambda(x) = \exp\big((\lambda \cdot T)(x) - G(\lambda)\big)\, q_0(x)$$

Properties of log partition function $G(\lambda)$

- Convex!
  - Proof via Hölder's inequality
- Strictly convex iff $T_1, \dots, T_n$ are **affinely independent** (on q_0's support)
  - **Affine independence**: $\lambda_1 T_1 + \cdots + \lambda_n T_n$ is constant iff $\lambda_1 = \cdots = \lambda_n = 0$
  - Proof via equality case of Hölder's inequality
- Gradient of $G(\lambda)$ w.r.t. λ:

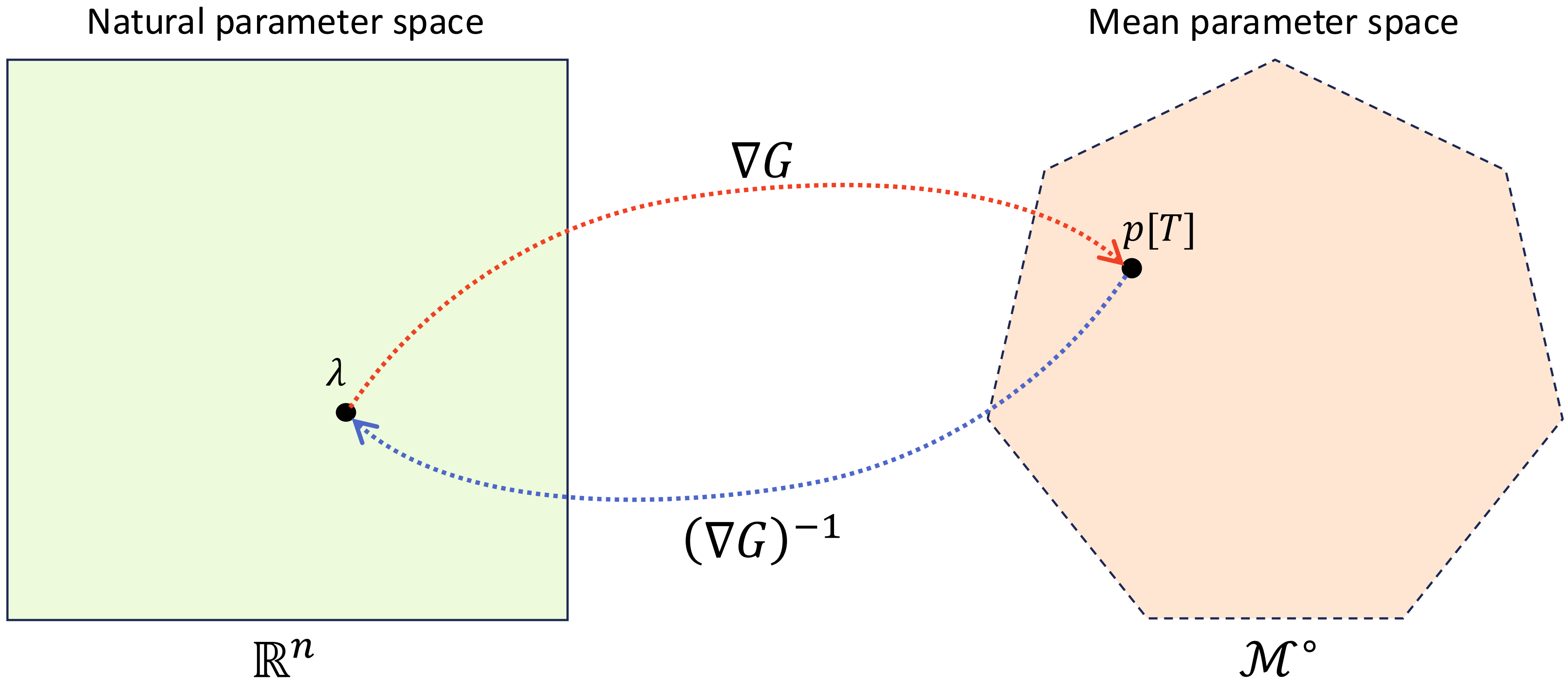
$$\nabla G(\lambda) = \frac{1}{Z(\lambda)} \sum_{x \in \mathcal{X}} T(x) \exp\big((\lambda \cdot T)(x)\big)\, q_0(x)$$

$$= \sum_{x \in \mathcal{X}} T(x) p_\lambda(x) = p_\lambda[T]$$

- Note: If G is strictly convex, then ∇G is 1-to-1!

The link between parameter spaces

Theorem: ∇G is 1-to-1 and $\nabla G(\mathbb{R}^n) = \mathcal{M}^\circ \coloneqq \{p[T] : p \in \Delta\}^\circ$



Exclusion of boundary points

In previous theorem, boundary points of $\mathcal{M}$ are excluded

- Example: $\mathcal{X} = \{0,1\}, T(x) = x, q_0(x) = \frac{1}{2}$
- Suppose $b = 1$, which is a valid "mean parameter":

  $$p[T] = b$$

  for $p(0) = 0,\ p(1) = 1$
- Cannot realize $p_\lambda[T] = 1$ by a Gibbs distribution since

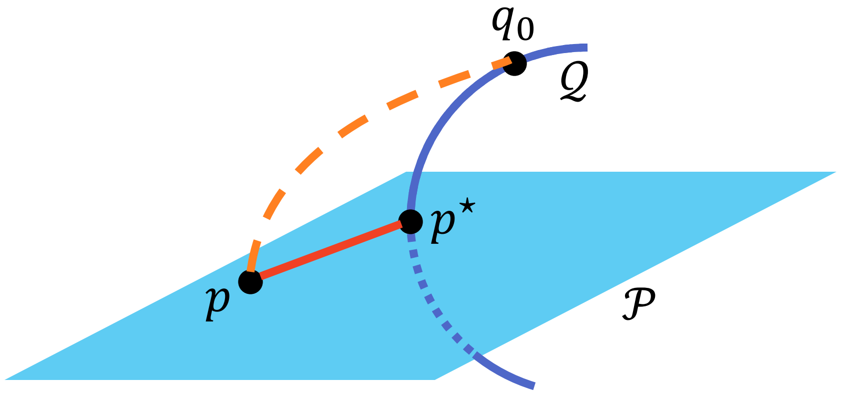
  $$p_\lambda(0) > 0$$

  for every $\lambda \in \mathbb{R}$ ☹

Information projection

Information projection

- Maxent solution also called **information projection** of q_0 onto $\mathcal{P}$
$$p^\star = \operatorname*{argmin}_{p \in \mathcal{P}} \mathrm{RE}(p, q_0)$$



- In fact, for any other $p \in \mathcal{P}$, we have a "Pythagorean identity"
$$\mathrm{RE}(p, q_0) = \mathrm{RE}(p, p^\star) + \mathrm{RE}(p^\star, q_0)$$

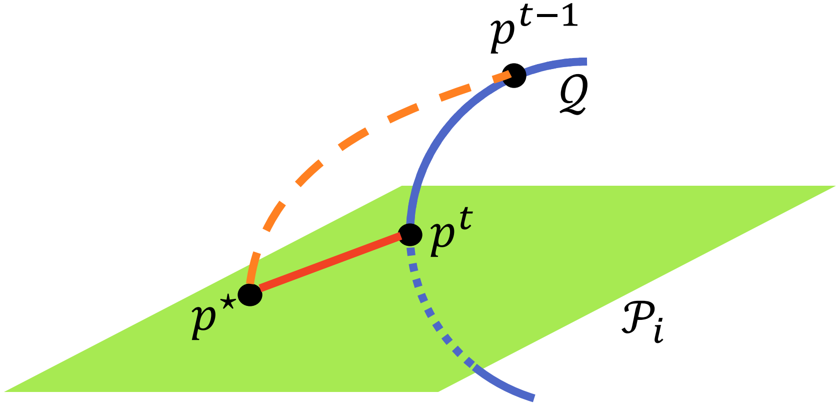
Proof of Pythagorean identity

For simplicity, assume $p^\star = p_\lambda \in \mathcal{Q}$ (a Gibbs distribution)

$$
\begin{aligned}
\mathrm{RE}(p, q_0) - \mathrm{RE}(p_\lambda, q_0) &= \mathrm{RE}(p, q_0) - p_\lambda\left[\log \frac{p_\lambda}{q_0}\right] \\
&= \mathrm{RE}(p, q_0) - p_\lambda[\lambda \cdot T - G(\lambda)] \\
&= \mathrm{RE}(p, q_0) - p[\lambda \cdot T - G(\lambda)] \\
&= \mathrm{RE}(p, q_0) - p\left[\log \frac{p_\lambda}{q_0}\right] \\
&= p\left[\log \frac{p}{q_0} - \log \frac{p_\lambda}{q_0}\right] \\
&= p\left[\log \frac{p}{p_\lambda}\right] = \mathrm{RE}(p, p_\lambda)
\end{aligned}
$$

Iterative projection algorithm

- Start with $p^0 = q_0$
- For $t = 1,2,\ldots$:
  - Pick some $i \in \{1,\ldots,n\}$, and let $\mathcal{P}_i = \{p \in \Delta : p[T_i] = b_i\}$
  - Let $p^t = \operatorname*{argmin}_{p \in \mathcal{P}_i} \mathrm{RE}(p, p^{t-1})$



- By Pythagorean identity,

$$\mathrm{RE}(p^\star, p^t) = \mathrm{RE}(p^\star, p^{t-1}) - \mathrm{RE}(p^t, p^{t-1})$$

Regularized maxent

Relaxing the expectation constraints

(Dudík, Phillips, Schapire, 2004)

- Suppose $b = \frac{1}{m}\sum_{j=1}^{m} T(x^j)$ for data set $x^1, \ldots, x^m \in \mathcal{X}$
- Even if $x^1, \ldots, x^m$ is i.i.d. sample from true information source p_{true}, we typically will not have $b = p_{\text{true}}[T]$, so doesn't make sense to require $p[T] = b$
- **Relaxed maxent problem**: Find $p \in \Delta$ that minimizes $\mathrm{RE}(p, q_0)$ while satisfying
$$|p[T_i] - b_i| \leq \beta_i \;\; \forall i = 1, \ldots, n$$
  - Regard $\beta_i \geq 0$ as "tuning parameters", based on deviation bounds for sample averages
- Dual objective (again, derived using method of Lagrange multipliers):

$$\sup_{\lambda \in \mathbb{R}^n} \underbrace{-\log q_0[\exp(\lambda \cdot T)] + \lambda \cdot b}_{\text{Original dual objective}} - \underbrace{\sum_{i=1}^{n} \beta_j |\lambda_j|}_{\text{Regularizer}}$$

Performance guarantee

- Pick any $\delta \in (0,1)$, and assume:
  - $T_i \colon \mathcal{X} \to [0,1]$ and $\beta_i = \beta \geq \sqrt{\log(2n/\delta)/(2m)}$ for all $i = 1, \ldots, n$
  - $x^1, \ldots, x^m$ is i.i.d. sample from p_{true}
  - $b_i = \frac{1}{m} \sum_{j=1}^{m} T_i\left(x^j\right)$ for all $i = 1, \ldots, n$
- With probability at least $1 - \delta$, solution to relaxed maxent problem $p_{\lambda^\star}$ satisfies

$$p_{\mathrm{true}}[\log p_{\lambda^\star}] \geq \sup_{\lambda \in \mathbb{R}^n} \left(p_{\mathrm{true}}[\log p_\lambda] - 2\|\lambda\|_1 \beta\right)$$