Notes on Generalizing the Maximum Entropy Principle to Uncertain Data

Kenneth Bogert*

May 2022

Abstract

The principle of maximum entropy is a broadly applicable technique for computing a distribution with the least amount of information possible constrained to match empirical data, for instance, feature expectations. We seek to generalize this principle to scenarios where the empirical feature expectations cannot be computed because the model variables are only partially observed, which introduces a dependency on the learned model. Generalizing the principle of latent maximum entropy[8], we introduce uncertain maximum entropy and describe an expectation-maximization based solution to approximately solve these problems. We show that our technique additionally generalizes the principle of maximum entropy and discuss a generally applicable regularization technique for adding error terms to feature expectation constraints in the event of limited data. We additionally discuss the use of black box classifiers with our technique, which simplifies the process of utilizing sparse, large data sets.

Keywords— entropy, maximum entropy, uncertainty, expectation maximization, partial observability, sparse data, deep learning, classifiers

1   Introduction

The principle of maximum entropy is a technique for finding a distribution over some given elements $X \in \mathbb{X}$ that contains the least amount of information in it while still matching some constraints. It has existed in various forms since the early 20th century but was formalized by Jaynes [4] in 1957. In its commonly encountered form, the constraints consist of matching sufficient statistics, or feature, expectations under the maximum entropy model being learned and those observed empirically.

However, in many cases the feature expectations are not directly observable. It could be the case that the model contains hidden variables, that some data is missing or corrupted by noise, or that X is only partially observable using some type of process or sensor.

As an example, let us take a simple natural language processing model. Using the principle of maximum entropy, each X will be a word in a vocabulary $\mathbb{X}$, and we wish to form a model that matches the empirical distribution of words in a given document, $\tilde{Pr}(X)$, according to the expectation of some interesting features $\phi_k(X)$.

However, if the data input into such a model is a voice recording then words are never directly observed. Instead, we may extract observations ω from the recording that only partially reveal the word being spoken, for instance, if ω corresponds to phonemes then $Pr(\omega|X)$, the probability of hearing a phoneme for a given word, will not be deterministic as different dialects and accents pronounce the same word in different ways. Further, a bad quality voice recording may cause

---

*kbogert@unca.edu, University of North Carolina Asheville

uncertainty in the phoneme being spoken, requiring the use of an even more general ω to correctly model the data.

With large amounts of sensor-produced data available and applications [2][1][6][5] that may make use of it providing the motivation, we seek to generalize the principle of maximum entropy to scenarios with partial observability of the modeled variables.

2 Background

2.1 Principle of Maximum Entropy

Commonly, the principle of maximum entropy is expressed as a non-linear program.

$$\max_{\Delta} \left(-\sum\nolimits_{X \in \mathbb{X}} Pr(X) \ log \ Pr(X) \right)$$

subject to

$$\sum\nolimits_{X \in \mathbb{X}} Pr(X) = 1$$
$$\sum\nolimits_{X \in \mathbb{X}} Pr(X)\phi_k(X) = \sum\nolimits_{X \in \mathbb{X}} \tilde{Pr}(X)\phi_k(X) \qquad \forall k \tag{1}$$

Notably, this program is known to be convex which provides a number of benefits. Particularly relevant is that we may find a close-form definition of $Pr(X)$, and solving the primal problem's dual is guaranteed to also solve the primal problem [3].

We begin by finding the Lagrangian relaxation of the program.

$$\mathcal{L}(\mathbb{X}, \lambda, \eta) = -\sum\nolimits_{X \in \mathbb{X}} Pr(X) \ log \ Pr(X) + \eta \left(\sum\nolimits_{X \in \mathbb{X}} Pr(X) - 1 \right) +$$
$$\sum_{k=1}^{K} \lambda_k \left(\sum_{X \in \mathbb{X}} Pr(X)\phi_k(X) - \sum_{X \in \mathbb{X}} \tilde{Pr}(X) \ \phi_k(X) \right) \tag{2}$$

Since the program is convex, the Lagrangian function must be as well. Therefore, when the Lagrangian's gradient is 0 we have found the global maximum. We now can find the definition of $Pr(X)$:

$$\frac{\partial \mathcal{L}(\mathbb{X}, \lambda, \eta)}{\partial Pr(X)} = -logPr(X) - 1 + \eta + \sum_{k=1}^{K} \lambda_k \phi_k(X)$$
$$0 = -logPr(X) - 1 + \eta + \sum_{k=1}^{K} \lambda_k \phi_k(X)$$
$$Pr(X) = \frac{e^{\sum_{k=1}^{K} \lambda_k \phi_k(X)}}{Z(\lambda)} \tag{3}$$

Where $Z(\lambda) = e^{-1}e^{\eta} = \sum_{X' \in \mathbb{X}} e^{\sum_{k=1}^{K} \lambda_k \phi_k(X')}$. Plugging our definition of $Pr(X)$ back into the Lagrangian, we arrive at the dual.

$$\mathcal{L}_{dual}(\lambda) = log \ Z(\lambda) - \sum_{k=1}^{K} \lambda_k \sum_{X \in \mathbb{X}} \tilde{Pr}(X)\phi_k(X) \tag{4}$$

Since the dual is necessarily convex, we find the gradient for use with gradient descent.

$$\frac{\partial \mathcal{L}_{dual}(\lambda)}{\partial \lambda_k} \;=\; \sum_X Pr(X)\phi_k(X) - \sum_X \tilde{Pr}(X)\phi_k(X)$$

(5)

Note that any convex optimization technique is a valid alternative to gradient descent.

2.2 Principle of Latent Maximum Entropy

First presented by Wang et al. [7], the principle of latent maximum entropy generalizes the principle of maximum entropy to models with hidden variables that are never empirically observed.

Split each X into Y and Z. Y is the component of X that is perfectly observed, Z is perfectly un-observed and completes Y. Thus, $X = Y \cup Z$ and $Pr(X) = Pr(Y, Z)$. Latent maximum entropy corrects for the hidden portion of X in the empirical data by summing over all $Z \in Z_Y$, which is every way of completing a given Y to arrive at a X.

$$\max_{\Delta} \left(-\sum_{X \in \mathbb{X}} Pr(X) \; log \; Pr(X) \right)$$

subject to

$$\sum_{X \in \mathbb{X}} Pr(X) = 1$$

$$\sum_{X \in \mathbb{X}} Pr(X)\phi_k(X) = \sum_{Y \in \mathbb{Y}} \tilde{Pr}(Y) \sum_{Z \in Z_Y} Pr(Z|Y)\phi_k(X) \qquad \forall k$$

(6)

Since $Pr(Z|Y)$ includes $Pr(X)$, the right side of the constraint contains a dependency on the model being learned, meaning the program is no longer convex and only an approximate solution can be found if we still desire a log-linear model for $Pr(X)$. This leads to an expectation-maximization approach to find a solution. To our knowledge Wang et al. 2001 [7] is the first to apply EM to the principle of maximum entropy to account for incomplete data.

The methodology and arguments used in this work is very similar to that used in Wang et al.'s [8] and so it will not be duplicated here. The reader is encouraged, however, to review Wang et al. [8] for more background and proofs.

3 Principle of Uncertain Maximum Entropy

Assume we want a maximum entropy model of some hidden variables $X \in \mathbb{X}$ given we have observations $\omega \in \Omega$. Critically, we desire that the model does NOT include ω as the observations themselves will pertain solely to the data gathering technique of the observing entity, not the elements or model being observed. We assume the existence of a static observation function $Pr(\omega|X)$. Our new non-linear program is:

$$\max_{\Delta} \left(-\sum_{X \in \mathbb{X}} Pr(X) \; log \; Pr(X) \right)$$

subject to

$$\sum_{X \in \mathbb{X}} Pr(X) = 1$$

$$\sum_{X \in \mathbb{X}} Pr(X)\phi_k(X) = \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_X Pr(X|\omega) \; \phi_k(X) \qquad \forall k$$

(7)

Notice that $Pr(X|\omega) = \frac{Pr(\omega|X)Pr(X)}{Pr(\omega)}$ and therefore in the infinite limit of data where $\tilde{Pr}(\omega) = Pr(\omega)$ the constraints are satisfied as:

3

$$\sum_{X \in \mathbb{X}} Pr(X)\phi_k(X) = \sum_{\omega \in \Omega} Pr(\omega) \sum_X Pr(X|\omega) \; \phi_k(X)$$

$$= \sum_{\omega \in \Omega} Pr(\omega) \sum_X \frac{Pr(\omega|X)Pr(X)}{Pr(\omega)} \; \phi_k(X)$$

$$= \sum_{\omega \in \Omega} \sum_X Pr(\omega|X)Pr(X) \; \phi_k(X)$$

$$= \sum_X Pr(X) \; \phi_k(X) \sum_{\omega \in \Omega} Pr(\omega|X)$$

$$= \sum_X Pr(X) \; \phi_k(X) \tag{8}$$

To attempt to solve Eq 7, we first take the Lagrangian.

$$\mathcal{L}(\mathbb{X}, \Omega, \lambda, \eta) \;=\; -\sum_{X \in \mathbb{X}} Pr(X) \; log \; Pr(X) + \eta \left(\sum_{X \in \mathbb{X}} Pr(X) - 1 \right) +$$
$$\sum_{k=1}^{K} \lambda_k \left(\sum_{X \in \mathbb{X}} Pr(X)\phi_k(X) - \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_X Pr(X|\omega) \; \phi_k(X) \right) \tag{9}$$

Now we find Lagrangian's gradient so that we can set it to zero and attempt to solve for $Pr(X)$.

$$\frac{\partial \mathcal{L}(\mathbb{X}, \Omega, \lambda, \eta)}{\partial Pr(X)} \;=\; -logPr(X) - 1 + \eta \;+$$
$$\sum_{k=1}^{K} \lambda_k \left(\phi_k(X) - \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \left(\phi_k(X) \frac{Pr(\omega|X)Pr(\omega) - Pr(\omega|X)^2 Pr(X)}{Pr(\omega)^2} \right) \right)$$
$$=\; -logPr(X) - 1 + \eta + \sum_{k=1}^{K} \lambda_k \phi_k(X)$$
$$-\sum_{k=1}^{K} \lambda_k \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \left(\phi_k(X) \frac{Pr(\omega|X)Pr(\omega) - Pr(\omega|X)^2 Pr(X)}{Pr(\omega)^2} \right) \tag{10}$$

Unfortunately, the existence of $Pr(X|\omega)$ on the right side of the constraints causes the derivative to be non-linear in $Pr(X)$. Instead, we will approximate $Pr(X)$ to be log-linear. In other words:

$$\frac{\partial \mathcal{L}(\mathbb{X}, \Omega, \lambda, \eta)}{\partial Pr(X)} \;\approx\; -logPr(X) - 1 + \eta + \sum_{k=1}^{K} \lambda_k \phi_k(X)$$

$$0 \;\approx\; -logPr(X) - 1 + \eta + \sum_{k=1}^{K} \lambda_k \phi_k(X)$$

$$Pr(X) \;\approx\; \frac{e^{\sum_{k=1}^{K} \lambda_k \phi_k(X)}}{Z(\lambda)} \tag{11}$$

Now we plug our approximation back into the Lagrangian to arrive at an approximate Dual:

$$\mathcal{L}_{dual}(\lambda) \;\approx\; log \; Z(\lambda) - \sum_{k=1}^{K} \lambda_k \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_X Pr(X|\omega)\phi_k(X) \tag{12}$$

We would now try to find the dual's gradient and use it to minimize the dual. Unfortunately the presence of $Pr(X|\omega)$ still admits no closed form solution in general. We will instead have to employ another technique to minimize it.

3.1 Expectation Maximization

Start with the log likelihood of all the observations:

$$
\begin{aligned}
L(\lambda) &= \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \, log \, Pr_\lambda(\omega) \\
&= \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \, log \, \sum_{X \in \mathbb{X}} Pr_\lambda(\omega, X) \\
&= \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \, log \, \sum_{X \in \mathbb{X}} \frac{Pr_\lambda(\omega, X)}{Pr_{\lambda'}(X|\omega)} Pr_{\lambda'}(X|\omega) \\
&\geq \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_{X \in \mathbb{X}} Pr_{\lambda'}(X|\omega) \, log \, \frac{Pr_\lambda(\omega, X)}{Pr_{\lambda'}(X|\omega)} \\
&= \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_{X \in \mathbb{X}} Pr_{\lambda'}(X|\omega) \, log \, Pr_\lambda(\omega, X) - \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_{X \in \mathbb{X}} Pr_{\lambda'}(X|\omega) \, log \, Pr_{\lambda'}(X|\omega) \\
&= \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_{X \in \mathbb{X}} Pr_{\lambda'}(X|\omega) \, log \, Pr_\lambda(\omega|X) Pr_\lambda(X) + H(\lambda') \\
&= \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_{X \in \mathbb{X}} Pr_{\lambda'}(X|\omega) \, log \, Pr_\lambda(\omega|X) + \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_{X \in \mathbb{X}} Pr_{\lambda'}(X|\omega) \, log \, Pr_\lambda(X) + H(\lambda') \\
&= \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_{X \in \mathbb{X}} Pr_{\lambda'}(X|\omega) \, log \, Pr(\omega|X) + Q(\lambda, \lambda') + H(\lambda') \qquad (13) \\
&= U^*(\lambda') + Q(\lambda, \lambda') + H(\lambda') \qquad (14)
\end{aligned}
$$

Eq 13 follows because $Pr(\omega|X)$ is the observation function which does not depend upon λ. This leaves $Q(\lambda, \lambda')$ as the only function which depends upon λ. The EM algorithm proceeds by maximizing Q, and upon convergence $\lambda = \lambda'$, at which time the likelihood of the data is at a local maximum.

$H(\lambda')$ is the conditional entropy on the latent variables, and $U^*(\lambda')$ is the expected log observations, which due the the observations not being included in the model only impacts the overall data likelihood, but not the model solution.

We now plug in a log-linear model for $Pr(X)$ to $Q(\lambda, \lambda')$:

$$
\begin{aligned}
Q(\lambda, \lambda') &= \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_{X \in \mathbb{X}} Pr_{\lambda'}(X|\omega) \, log \, Pr_\lambda(X) \\
&= \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_{X \in \mathbb{X}} Pr_{\lambda'}(X|\omega) \left(\sum_{k=1}^{K} \lambda_k \phi_k(X) - \, log \, Z(\lambda) \right) \\
&= -log \, Z(\lambda) + \sum_{k=1}^{K} \lambda_k \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_{X \in \mathbb{X}} Pr_{\lambda'}(X|\omega) \phi_k(X) \qquad (15)
\end{aligned}
$$

Notice that Eq. 15 is similar to Eq. 12. One important difference is that Eq. 15 is easier to solve, as $Pr(X|\omega)$ depends on λ' and not λ. In fact, maximizing $Q(\lambda, \lambda')$ is equivalent to solving the following program:

$$\max_{\Delta} \left(- \sum\nolimits_{X \in \mathbb{X}} Pr_\lambda(X) \ log \ Pr_\lambda(X) \right)$$

subject to

$$\sum\nolimits_{X \in \mathbb{X}} Pr_\lambda(X) = 1$$

$$\sum\nolimits_{X \in \mathbb{X}} Pr_\lambda(X) \phi_k(X) = \sum\nolimits_{\omega \in \Omega} \tilde{Pr}(\omega) \sum\nolimits_X Pr_{\lambda'}(X|\omega) \ \phi_k(X) \qquad \forall k \qquad (16)$$

which equals Eq.7 at convergence. We now arrive at the following Expectation-Maximization algorithm:

Initial Start: Randomly initialize λ'
E Step: Using λ', compute $\hat{\phi}_k = \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_X Pr_{\lambda'}(X|\omega) \ \phi_k(X)$
M Step: Solve the following convex program to arrive at a new λ:

$$\max_{\Delta} \left(- \sum\nolimits_{X \in \mathbb{X}} Pr_\lambda(X) \ log \ Pr_\lambda(X) \right)$$

subject to

$$\sum\nolimits_{X \in \mathbb{X}} Pr_\lambda(X) = 1$$

$$\sum\nolimits_{X \in \mathbb{X}} Pr_\lambda(X) \phi_k(X) = \hat{\phi}_k \qquad \forall k \qquad (17)$$

Then set $\lambda' = \lambda$
Repeat: Until λ converges

4 Specializations

Here we demonstrate that the principle of uncertain maximum entropy generalizes both the principle of maximum entropy and the principle of latent maximum entropy[8] by showing that we recover these earlier methods when certain specific conditions are met.

4.1 Principle of Maximum Entropy:

We recover the Principle of Maximum Entropy if $Pr(X|\omega) \in \{0, 1\} \ \forall \ X, \omega$ and $\exists \ \omega \ni \ Pr(X|\omega) = 1 \ \forall X$. In other words, each ω specifies a single X deterministically. Note that the reverse is not necessarily true, $Pr(\omega|X)$ need only be deterministic if $|\Omega| = |\mathbb{X}|$. However, for a given X specified by a given ω:

$$Pr(X|\omega) = \frac{Pr(\omega|X)Pr(X)}{Pr(\omega)}$$

$$1 = \frac{Pr(\omega|X)Pr(X)}{Pr(\omega)}$$

$$Pr(\omega) = Pr(\omega|X)Pr(X)$$

Therefore, in Eq 10 (the Lagrangian's gradient), the final term is always zero and we find $Pr(X)$ is log linear (without approximation), and Eq. 12 is exact. Furthermore, as $Pr(X|\omega)$ is unaffected by λ the gradient of Eq. 12 may now be found as:

$$\frac{\partial \mathcal{L}_{dual}(\lambda)}{\partial \lambda_k} = \sum_X Pr(X)\phi_k(X) - \sum_{\omega \in \Omega} \tilde{Pr}(\omega) \sum_X Pr(X|\omega)\phi_k(X) \qquad (18)$$

Thus, we do not need to use EM to solve this problem, and we have arrived at a principle of Maximum Entropy solution. To see terms that exactly match, let $|\Omega| = |\mathbb{X}|$, then $\tilde{Pr}(\Omega) = \tilde{Pr}(X)$ and $\sum_X Pr(X|\omega) = \sum_{X'} Pr(X'|X)$.

$$\begin{aligned}
\mathcal{L}_{dual}(\lambda) &= logZ(\lambda) - \sum_{k=1}^{K} \lambda_k \sum_{X \in \mathbb{X}} \tilde{Pr}(X) \sum_{X'} Pr(X'|X)\phi_k(X) \\
&= logZ(\lambda) - \sum_{k=1}^{K} \lambda_k \sum_{X \in \mathbb{X}} \tilde{Pr}(X)\phi_k(X)
\end{aligned}$$
$$(19)$$

4.2 Principle of Latent Maximum Entropy:

[8] This technique breaks up X into two components, Y which is perfectly observed, and Z which is missing (perfectly un-observed) and $X = Y \cup Z$. To show that Maximum Entropy with Uncertain Observations generalizes latent maximum entropy, we must show a reduction of the right side of the main constraint to $\sum_Y \tilde{Pr}(Y) \sum_{Z \in Z_Y} Pr(Z|Y)\phi_k(X)$ when $Pr(Y|\omega) \in \{0,1\} \; \forall \; Y, \omega$ and $\exists \; \omega \ni Pr(Y|\omega) = 1 \; \forall \; Y$. In other words, each ω specifies a single Y deterministically. Note that the reverse is not necessarily true, $Pr(\omega|Y)$ need only be deterministic if $|\Omega| = |\mathbb{Y}|$.

Using this definition,

$$\begin{aligned}
Pr(Y|\omega) &= \frac{Pr(\omega|Y)Pr(Y)}{Pr(\omega)} \\
1 &= \frac{Pr(\omega|Y)Pr(Y)}{Pr(\omega)} \\
Pr(\omega) &= Pr(\omega|Y)Pr(Y)
\end{aligned}$$

Now note that $Pr(X) = Pr(Y, Z)$, we have

$$\begin{aligned}
Pr(X|\omega) &= \frac{Pr(\omega|Y, Z)Pr(Y, Z)}{Pr(\omega)} \\
&= \frac{Pr(\omega|Y, Z)Pr(Z|Y)Pr(Y)}{Pr(\omega|Y)Pr(Y)} \\
&= \frac{Pr(\omega|Y, Z)Pr(Z|Y)}{Pr(\omega|Y)} \\
&= \frac{Pr(\omega|Y)Pr(Z|Y)}{Pr(\omega|Y)} \qquad (20) \\
&= Pr(Z|Y) \qquad (21)
\end{aligned}$$

Since Z is perfectly unobserved, $Pr(\omega|Y, Z) = Pr(\omega|Y)$ on eq 20. To match terms exactly, let $|\Omega| = |Y|$, then $\tilde{Pr}(\Omega) = \tilde{Pr}(Y)$. Notice, whenever $Pr(Y|\omega) = 0, Pr(X|\omega) = 0$. Therefore we may ignore the summation term in these cases, and only consider $Z \in Z_Y$:
$$\sum_{\omega} \tilde{Pr}(\omega) \sum_X Pr(Z|Y)\phi_k(X) = \sum_Y \tilde{Pr}(Y) \sum_{Z \in Z_Y} Pr(Z|Y)\phi_k(X)$$

5 Large, sparse observation sets

The desire to automate inference has driven the use of extremely large, sparse datasets produced by machine sensors as the input into various learning models. Often techniques such as deep neural networks are used to transform the sparse dataset into dense data, perhaps the model elements directly. These techniques may be trained by making use of supervised learning on a subset of the available data that has been manually labeled. As the output learned model may be a black box, no human discernible observation features may be available for examination for use in uMaxEnt. Here we extend the principle of uncertain maximum entropy to these black box scenarios.

Suppose we have an enormous, sparse dataset $\mathbb{R}$ from which samples r are produced from the model element's true observation features, ie. $Pr(r|\omega)$ (r stands for *raw data*). These r samples are what is received by the observer, as Ω is unknown to the observer, and may be thought of as encoded (possibly partially) into r. For example, if the observer is using a RGB camera, ω may be a 3D mesh describing the full visual representation of a particular X and r is a 2D RGB image of the mesh.

Suppose we are given a set of samples from $\mathbb{R}$ labeled by a human. To increase generality, we allow the labels to be from a different, though related, set than $\mathbb{X}$. Let $\xi \in \Xi$ be these labels and define a function $d(X, \xi) \to \{0, 1\}$ that is 1 when a given X maps to a given ξ. For simplicity of argument we restrict d such that each X maps to only one ξ deterministically. Note that the opposite need not be true, one ξ may map probabilistically to many X. Extension to more general configurations is straightforward and only involves modifying $Pr(\xi|X)$ appropriately, we will not discuss this further here.

Now, we may employ some method to classify all received r into ξ. Let $F(r) \to \xi$ be the function learned by this method, and let us further assume this method comes with statistical performance metrics such as precision and recall. Then, we use F to classify all available sparse data into $\tilde{Pr}(\xi)$ and our new uMaxEnt constraints for this scenario are:

$$\sum_{X \in \mathbb{X}} Pr(X)\phi_k(X) = \sum_{\xi \in \Xi} \tilde{Pr}(\xi) \sum_X Pr(X|\xi)\phi_k(X) \tag{22}$$

Where $Pr(X|\xi) = \frac{Pr(\xi|X)Pr(X)}{Pr(\xi)}$, and $Pr(\xi|X)$ is the probability that F outputs ξ when the true, underlying model element present is X. This is given by the method's performance metrics, though possibly with appropriate modification to account for the difference between Ξ and $\mathbb{X}$. Note that in the event the classification method used is perfect, these new constraints revert to either latent maximum entropy (when $\Xi \subset \mathbb{X}$) or standard maximum entropy (when $\Xi = \mathbb{X}$).

5.1 Uncertain classification

Suppose that the classification method used cannot be certain as to which ξ should be output for a given r and instead produces a distribution over ξ, $Pr(\xi|r)$. This provides only partial information of which ξ is present, but has an advantage in that the method encodes the accuracy of its output into the output distribution itself. This in turn greatly simplifies $Pr(\xi|X) = d(X, \xi)$.

Our first attempt at using this distribution may be to find the expected ξ as follows:

$$\sum_{X \in \mathbb{X}} Pr(X)\phi_k(X) = \sum_r \tilde{Pr}(r) \sum_{\xi \in \Xi} Pr(\xi|r) \sum_X Pr(X|\xi)\phi_k(X) \tag{23}$$

However, this faces an issue as the distribution $Pr(\xi|r)$ is produced using the training data set, and not the specific dataset under consideration. To see this, suppose the method used is parameterized with θ, and we provide an infinite amount of data such that $\tilde{Pr}(r) = Pr(r)$:

$$\sum_r Pr(r) \sum_{\xi \in \Xi} Pr_\theta(\xi|r) \sum_X Pr(X|\xi)\phi_k(X)$$

$$= \sum_r \sum_{\xi \in \Xi} Pr_\theta(\xi, r) \sum_X Pr(X|\xi)\phi_k(X)$$

$$= \sum_{\xi \in \Xi} Pr_\theta(\xi) \sum_X Pr(X|\xi)\phi_k(X)$$

$$\neq \sum_{\xi \in \Xi} \sum_X Pr(X, \xi)\phi_k(X) \tag{24}$$

$$\tag{25}$$

Because the **training** distribution over ξ, $Pr_\theta(\xi)$, can vary dramatically from the target distribution $Pr(\xi)$, we cannot guarantee that this method produces an effective approximation. For instance, suppose in the training set the distribution over ξ was deliberately chosen to be uniform in order to prevent bias in the learning, whereas this distribution is highly unlikely to be the correct one in an inference task.

To correct for this, we examine $Pr_\theta(\xi|r)$ using Baye's law. Note that even if the method used does not allow for these components to be separated as shown they still must be represented in some capacity in order to produce a valid distribution.

$$Pr_\theta(\xi|r) = \frac{Pr_\theta(r|\xi)Pr_\theta(\xi)}{Pr_\theta(r)} \tag{26}$$

We note that $Pr_\theta(r)$ is a normalizer, and so we target $Pr_\theta(\xi)$ and replace it with $Pr(\xi)$, then normalize to obtain an updated distribution.

$$Pr_\theta(\xi|r)\frac{Pr(\xi)}{Pr_\theta(\xi)} = \frac{\nu \; Pr_\theta(r|\xi)Pr_\theta(\xi)Pr(\xi)}{Pr_\theta(\xi)}$$

$$= \nu' \; Pr_\theta(r|\xi)Pr(\xi)$$

$$= \frac{Pr_\theta(r|\xi)Pr(\xi)}{\sum_{\xi'} Pr_\theta(r|\xi')Pr(\xi')} \tag{27}$$

Where ν and ν' are normalizers, and differ as we require renormalization after the correction. Now, notice $Pr(r) = \sum_\xi Pr(r|\xi)Pr(\xi)$, which differs from the normalizer above only in the term $Pr_\theta(r|\xi)$, which is the observation model being learned by the classification technique. This term is expected to approximate the true observation model as closely as possible, as that is the whole purpose of employing the technique!

So we arrive at, in the case of infinite data:

$$\sum_{X \in \mathbb{X}} Pr(X)\phi_k(X) = \sum_r Pr(r) \sum_{\xi \in \Xi} \frac{Pr_\theta(\xi|r)Pr(\xi)}{Pr_\theta(\xi)} \sum_X Pr(X|\xi)\phi_k(X)$$

$$\approx \sum_r \sum_{\xi \in \Xi} Pr_\theta(r|\xi)Pr(\xi) \sum_X Pr(X|\xi)\phi_k(X)$$

$$= \sum_{\xi \in \Xi} Pr(\xi) \sum_X Pr(X|\xi)\phi_k(X) \sum_r Pr_\theta(r|\xi)$$

$$= \sum_{\xi \in \Xi} Pr(\xi) \sum_X Pr(X|\xi)\phi_k(X)$$

$$= \sum_{\xi \in \Xi} \sum_X Pr(X,\xi)\phi_k(X)$$

$$= \sum_X Pr(X)\phi_k(X) \tag{28}$$

The quality of the approximation is now controlled by the quality of the classification technique, this a desirable trait as the classification technique's quality is controlled by the engineers building or training it.

This variant of uMaxEnt incorporates $Pr(X)$ twice since $Pr(\xi) = \sum_X Pr(\xi|X)Pr(X)$. Notice that even in the event that $\Xi = \mathbb{X}$ we still have a uMaxEnt problem, due to the presence of this second $Pr(X)$ and ultimately caused by the uncertainty in $Pr_\theta(\xi|r)$.

6 Discussion

The principle of uncertain maximum entropy makes explicit that the choice of model influences results by including $Pr(X)$ in the empirical side of the constraints. In cases where uncertainty exists in $Pr(X|\omega)$ this technique ensures a model is found that is consistent with the available information and not over-committed to the specific observations received, as would be the case with ignoring the uncertainty and using the principle of maximum entropy, perhaps by taking the expectation, mean, or maximum X given ω.

Another benefit of this technique is existing $Pr(X)$ priors may be used in the first E step of the expectation-maximization algorithm, somewhat similar to how it is done with Bayesian methods, as opposed to uninformative priors. This can help bias the results to reflect earlier experiences that cannot, for whatever reason, be included in $\tilde{Pr}(\omega)$.

References

[1] K. Bogert and P. Doshi. A hierarchical bayesian process for inverse rl in partially-controlled environments. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 145–153, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems.

[2] K. Bogert, J. F.-S. Lin, P. Doshi, and D. Kulic. Expectation-maximization for inverse reinforcement learning with hidden data. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 1034–1042, 2016.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. 2002.

[4] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[5] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert. Activity forecasting. *Computer Vision–ECCV*, pages 201–214, 2012.

[6] S. Shahryari and P. Doshi. Inverse reinforcement learning under noisy observations. *arXiv preprint arXiv:1710.10116*, 2017.

[7] S. Wang, R. Rosenfeld, and Y. Zhao. Latent maximum entropy principle for statistical language modeling. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01.*, pages 182–185. IEEE, 2001.

[8] S. Wang, D. Schuurmans, and Y. Zhao. The latent maximum entropy principle. *ACM Trans. Knowl. Discov. Data*, 6(2), July 2012.