Minimizing Free Energy of Stochastic Functions of Markov Chains

Rita Singh

Carnegie Mellon University, Pittsburgh PA 15213, USA,
rsingh@cs.cmu.edu,
WWW home page: http://cs.cmu.edu/~rsingh

Abstract. Automatic speech recognition has generally been treated as a problem of Bayesian classification. This is suboptimal when the distributions of the training data do not match those of the test data to be recognized. In this paper we propose an alternate analogous classification paradigm, in which classes are modeled by thermodynamic systems, and classification is performed through a *minimum energy* rule. Bayesian classification is shown to be a specific instance of this paradigm when the temperature of the systems is unity. Classification at elevated temperatures naturally provides a mechanism for dealing with statistical variations between test and training data.

Keywords: Bayes classification, free energy, temperature, speech recognition

1 Introduction

In the usual rule for Bayesian classification any data X is assigned to the class that is most likely to have generated it. Formally, if we represent the class assignment of X as $c(X)$, the classification rule is given by

$$c(X) = \arg\max_C P(C)P(X|C) \tag{1}$$

where C represents any class, $P(C)$ is the *a priori* probability of C, and $P(X|C)$ is the probability distribution of data from class C. In the context of automatic speech recognition, the classes are actually word sequences [1]. The Bayes classification rule attempts to identify the *a posteriori* most likely word sequence, given (features derived from) a recording X.

The Bayes classification rule is optimal when $P(X|C)$ is the *true* class-conditional probability distribution of data for C. In practical scenarios, however, the true distribution $P(X|C)$ is not known and must be approximated by a *model* $\hat{P}(X; \Lambda_C)$, the parameters Λ_C of which must be learned from training data. The model $\hat{P}(X; \Lambda_C)$ is generally learned to be close (in the sense of KL divergence) to the distribution of the training data, and often does not adequately model the *test* data. As as result, the classification rule of (1) is suboptimal and can result in significantly degraded performance when used in a speech recognizer.

In [2, 3] we have proposed an alternative formalism for classification in such scenarios. Instead of assuming that $\hat{P}(X; \Lambda_C)$ represents a class-specific probability distribution, we interpret it as a *thermodynamic system*, which has resulted in an observation X. Subsequently, we replace the "maximum probability" criterion deriving from the stochastic-process interpretation which leads to Bayes classification rule, with a "minimum energy" criterion: the observation X is now assigned to the class C whose system must expend the least energy to generate it. Designating the energy as $F_C(X)$, the modified classification rule is

$$c(X) = \arg\min_C F_C(X) \tag{2}$$

The distinction between the probability-based rule of (1), and (2) can be resolved by defining $F_C(X) = -\log P(C, X; \Lambda_C)$. Indeed, such an equivalence is commonly ascribed, and has been drawn upon in the definition of stochastic models such as the Gibbs distribution [4], or even the normal distribution, where the log probability is analogous to common definitions of energy in a data or vector [5].

Thermodynamic systems, however, also include a *temperature* parameter. In the physical world the temperature of the system characterizes the fluctuation of state of the system, effectively characterizing the variation in any measurements of it – the greater the temperature the greater the variation will be. In our classification framework, the temperature parameter may be analogously considered as characterizing the increased variation in observations. At the specific setting of $T = 1$, the probabilistic and energy classification rules become identical; at higher values however, the energy-based mechanism naturally allows for greater variation in the data, such as the differences between training and test data.

In the subsequent sections we will first describe the general Themodynamic principle of free energy (Sect. 2), followed by a brief outline of minimum-free energy classification (Sect. 3) and how it applies to automatic speech recognition (Sect. 4). We then present experimental evidence of the effectiveness of the formulation (Sect. 5) and discussion (Sect. 6).

2 Free Energy of a Stochastic System

A thermodynamic system at temperature T can exist in one of a large (potentially infinite) number of states [6]. At each state s the system has an energy E_s. If the probability of state s is given by $P_T(s)$, the *internal energy* of the system, representing the capacity of the system to do work, is given by the average: $U_T = \sum_s P_T(s) E_s$. This capacity is counteracted by its internal disorder, which is factored into its entropy $H_T = -\sum_s P_T(s) \log P_T(s)$ and the temperature T of the system. The *Helmholtz free energy* of the system, measuring the useful work obtainable from the system when it is closed, is thus defined by

$$F_T = U_T - TH_T = \sum_s P_T(s) E_s + T \sum_s P_T(s) \log P_T(s) \tag{3}$$

At constant temperature, systems will drift towards the lowest free-energy states [6], adjusting probabilities $P_T(s)$ until F_T is minimized. The distribution $P_T(s)$ at *thermal equilibrium*, obtained by minimizing F_T, is the Gibbs distribution

$$P_T(s) = \frac{1}{Z} \exp\left(\frac{-E_s}{T}\right) \tag{4}$$

where Z is a normalizing term. The corresponding *equilibrium free energy* is

$$F_T = -T \log \sum_s \exp\left(\frac{-E_s}{T}\right) \tag{5}$$

3 Classification with Free Energy

Consider a class with a stochastic generative latent-variable model that assigns a probability $P(X|C) = \sum_s P(s|C)P(X|C,s)$ to any observation. To generate any observation, the generative process must be in any latent state s and draw an observation from the state-conditioned distribution $P(X|C,s)$.

For energy-based classification we model every class C instead by a thermodynamic system that can exist in one of a set $\mathcal{S}_C$ of states. Within any state s the system must have an energy $E_s^C(X)$ to result in the observation X. The equilibrium free energy of this system, when it is at temperature T, is hence given by

$$F_T^C(X) = -T \log \sum_s \exp\left(\frac{-E_s^C(X)}{T}\right) \tag{6}$$

The "energy" of each state is equated to negative log-likelihood of the combination of the state and the observation, $E_s^C(X) = -\log P(X,s,C)$ – intuitively, the greater the energy needed to exhibit X, the less likely the system is to visit the corresponding state. Using these values, the free energy of the system for any class comes out as

$$F_T^C(X) = -\log P(C) - T \log \sum_s \exp\left(\frac{\log P(X,s|C)}{T}\right) \tag{7}$$

We specify the minimum-energy classification rule as follows: the observation X is assigned to the class that has the lowest free energy for X.

$$c(X) = \arg\min_C F_T^C(X) = \arg\min_C \left(-\log P(C) - T \log \sum_s P(X,s|C)^{\frac{1}{T}}\right) \tag{8}$$

This is a natural extension of the principle that thermodynamic systems evolve towards minimum-energy configurations. Note that the objective in (8) remains a function of the temperature parameter T. As T increases and the internal disorder in the systems increases, the systems for the various classes will more frequently visit low-energy states associated with X; in the limit T dominates

and all classes are equally capable of generating observation X. From a classification perspective, T characterizes external influences such as noise or other factors that increase the entropy of the systems. Note that at $T = 1$ (the "quiescent" condition) (8) reduces to a conventional Bayesian classifier of (1).

4 Minimum Free Energy Decoding with Hidden Markov Models

A particularly interesting family of stochastic models that can be cast into the free-energy framework are stochastic functions of Markov chains, also known as Hidden Markov Models (HMMs) [7]. HMMs are frequently employed in automatic speech recognition systems [1]. HMM-based speech recognition systems formulate the Bayes classification paradigm as identifying the word sequence with the *a posteriori* most likely state sequence for any speech recording X [8].

$$\hat{W} = \arg\max_{W} \max_{S} P(W)P(S|W)P(X|S,W) \tag{9}$$

where W represents any word sequence, $P(W)$ is the *a priori* probability of W, S is a state sequence through the HMM for W, and $P(S|W)$ represents its probability. The state output distributions of the HMM are often modeled by mixture distributions, typically Gaussian mixtures. Thus the classification equation can be re-written as

$$\begin{aligned}\hat{W} &= \arg\max_{W} \max_{S} P(W,S) \prod_{t=1}^{T} P(X_t|s_t) \\ &= \arg\max_{W} \max_{S} P(W,S) \prod_{t=1}^{T} \sum_{k} P(k|s_t) N(X_t; \mu_{s_t,k}, \Sigma_{s_t,k})\end{aligned} \tag{10}$$

Here X_t is the t^{th} vector in X and s_t is the t^{th} state in S. $N()$ represents a Gaussian distribution, $P(k|s_t)$ represents the mixture weight of the k^{th} Gaussian in the Gaussian mixture distribution for state s_t, and $\mu_{s_t,k}$ and $\Sigma_{s_t,k}$ represent the mean and covariance of the k^{th} Gaussian in s_t.

We can define $K = k_1, k_2, \ldots, k_T$, representing a sequence of Gaussians, one each from states s_1-s_T. If each state is represented by a mixture of M Gaussians, there are M^T such sequences. Combining W and S into a single variable $\mathcal{W}$, (10) can be rewritten as:

$$\hat{\mathcal{W}} = \arg\max_{\mathcal{W}} \log P(\mathcal{W}) + \log \sum_{K} P(X,K|\mathcal{W}) \tag{11}$$

where

$$P(X,K|\mathcal{W}) = \prod_{t=1}^{T} P(k_t|s_t) N(X_t; \mu_{s_t,k}, \Sigma_{s_t,k}) \tag{12}$$

The above is now easily recast into minimum-energy classification. Each class $\mathcal{W}$ is represented by a thermodynamic system, which can be in one of a M^T states,

where each state is a Gaussian sequence K. The energy of any state is given by $E_K^{\mathcal{W}} = -\log P(X, K, \mathcal{W})$. Consequently, the minimum free-energy classification rule of (7) becomes, with minimal manipulation,

$$\hat{\mathcal{W}} = \arg\min_{\mathcal{W}} -\log P(\mathcal{W}) - T \sum_t \log \sum_k P(k_t|s_t)^{\frac{1}{T}} N(X_t; \mu_{s_t,k}, \Sigma_{s_t,k})^{\frac{1}{T}} \quad (13)$$

This modified classification rule requires only *minimal* changes to the conventional Viterbi decoder. The computation of state output distribution values as $\sum_k P(k_t|s_t)N(X_t; \mu_{s_t,k}, \Sigma_{s_t,k})$ is replaced by $\left(\sum_k P(k_t|s_t)^{\frac{1}{T}} N(X_t; \mu_{s_t,k}, \Sigma_{s_t,k})^{\frac{1}{T}}\right)^T$. The rest of the decoder remains unchanged. We refer to this modified decoding strategy as "minimum-energy decoding".

5 Experiments

We expect the benefits of minimum-energy recognition to be exhibited primarily when there is mismatch between the distributions employed by the recognizer and the test data. Our experiments were therefore aimed at evaluating the effect of minimum-energy decoding under conditions of mismatch. One of the most common reasons for mismatch in speech recognition systems is noise: test data to be recognized will frequently be corrupted by various types of noise not seen in the training data. Note that noise robustness is not the focus of this paper; rather it is the mismatch between the acoustic models and the test data.

Table 1. Performance of minimum-energy speech recognition in terms of word error rate (%). The T=1.0 column corresponds to conventional decoding, representing Bayesian classification. The bold numbers are the best results obtained in each row.

Temp	*1.0*	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0dB	*92.7*	90.8	87.4	85.4	79.9	79.3	78.2	77.9	**75.8**	77.8	82.3
5dB	*65.3*	62.4	57.4	55.8	53.4	51.4	50.5	49.2	**48.3**	51.3	57.7
10dB	*47.6*	46.9	45.1	44.8	42.8	38.2	37.4	**36.6**	37.8	41.2	47.2
15dB	*36.2*	36.1	35.1	33.5	31.9	**30.2**	30.8	31.8	34.2	37.2	40.1
20dB	*27.2*	26.8	25.1	**24.6**	24.6	25.4	27.2	29.4	32.4	35.2	38.1

We conducted experiments on the Fisher database [9] digitally corrupted by noise to introduce mismatch. The training data comprised the Fisher Phase I corpus (LDC catalog No. LDC2004S13), including 5,850 two-channel audio files, each containing a full conversation of up to 10 minutes. 111157 speech segments from the corpus, representing nearly the entire data, minus our held out test set, were used to train the models. A set of 10,000 segments from the same data were used as our held-out designated test set. The test set were corrupted to various signal-to-noise ratios (SNR) by babble noise to introduce mismatch with respect to the the training set.

We used the Carnegie Mellon University's Sphinx-3 triphone-based automatic speech recognition system [10] to perform all our experiments. All models were 3-state left-to-right Bakis topology HMMs. A total of 5000 tied states, each modeled by a mixture of 16 Gaussians, were employed. The language model was trained from the Fisher training corpus and the Switchboard corpus. The baseline recognition word error rate on the uncorrupted test set was 14.3%.

The test data were recognized at several temperatures. Table 1 shows the word error rates obtained at each SNR, against the temperature at which the data were decoded. The column in the table corresponding to $T = 1.0$ is identical to the standard Bayesian decoding, as explained earlier.

We note from the results that the optimal recognition performance is *not* obtained at $T = 1$. The best result in all cases occurs at an elevated temperature. Moreover, as the SNR decreases and, consequently, the degree of mismatch between the training and test data increases, the optimal temperature increases. Thus, while the optimal temperature at 0dB is close to 2.0, at 20dB the optimal temperature is 1.3. At greater mismatch, *e.g.* at 0dB, the improvement from increased temperature is quite dramatic, amounting to about 17% absolute.

6 Conclusions

Elevation of temperature is observed to result in significantly improved recognition under conditions of mismatch. Considering that just a simple adjustment has been made to the manner in which state-output probabilities are computed during decoding in order to achieve this, the improved classification scheme is promising for use in speech recognizers. It must be noted that although these improvements are not as large as that improved with sophisticated noise compensation algorithms, that is not the objective of our solution. The proposed algorithm makes no assumptions about the *reason* for the mismatch; the only assumption is that while the systematic differences between classes persist in the test data, the actual distribution may be shifted with respect to the training data. Our purpose is to demonstrate that the proposed approach, which is a natural extension of conventional Bayesian classification, could be used to good effect under such conditions.

A key question that remains to be answered is that of *selecting* the optimal temperature in an unsupervised manner. We continue to explore this problem.

More generally, the notions of "temperature" and "free energy" have often been invoked in the context of annealing for optimization of objective functions defined over a continuous support [11]. Classification, on the other hand, is typically a search over a discrete support, and not usually viewed as an optimization problem. This is generally considered to be distinct from the situations where notions of free energy and temperature may be invoked. The novelty of our approach is to view the latter as a special case of optimization, where the task is to find the optimal value over a discrete support. In this context, automatic speech recognition systems present an interesting case – although the support remains discrete, it is inifinite, representing all possible sentences that may be spoken,

suggesting that the concept of annealing may be drawn upon if the search space could somehow be ordered and represented over a continuum. However, how this may be done is unclear, and this remains a topic for future research.

References

1. Singh, R.; Raj, B.; Virtanen, T.: *The Basics of Automatic Speech Recognition.* in: Techniques for Noise Robustness in Automatic Speech Recognition. John Wiley and Sons Inc. (2012)
2. Singh, R.: *Audio classification with thermodynamic criteria.* Proc. IEEE Int. Workshop on Cloud Computing for Signal Processing, Coding and Networking. (2014)
3. Singh, R.; Kumatani, K.: *Free energy for speech recognition.* Proc. Int. Conf. Acoustics, Speech and Signal Processing. (2015)
4. Landau, L. D.; Lifshitz, E. M.: *Statistical physics: Course of theoretical physics 5 (3 ed.).* Oxford: Pergamon Press. (1980)
5. Ranzato M. A.; Boureau Y. L.; Yann L. C.: *Sparse feature learning for deep belief networks.* Proc. Advances in Neural Information Processing Systems. 21, 1185–1192. (2008)
6. Callen, H. B.: *Thermodynamics and an introduction to thermostatistics.* John Wiley and Sons Inc. (1985)
7. Baum, L. E.; Petrie, T.: *Statistical inference for probabilistic functions of finite state Markov chains.* The Annals of Mathematical Statistics. 37(6), 1554–1563. (1966)
8. Huang, X.; Acero, A.; Hon, H.W.: *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice Hall (2001)
9. Cieri, C.; Miller, D.; Walker, K.: *The Fisher Corpus: a Resource for the next generations of speech-to-text.* Intl. Conf. on Language Resources and Evaluation. (2004)
10. Website: http://cmusphinx.sourceforge.net
11. Aarts, E.; Korst, J.: *Simulated annealing and Boltzmann machines.* John Wiley and Sons Inc. (1988)