

# Rate Distortion with Bregman Divergences

Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh and Srujana Merugu  
{abanerje,ghosh,inderjit,merugu}@ece.utexas.edu  
Department of Electrical and Computer Engineering  
University of Texas at Austin  
Austin, TX 78712

Several clustering algorithms have traditionally used techniques from the parametric mixture model learning literature. Recent years have seen a lot of interest and research on information theoretic formulations of clustering [7] that partially draw from Shannon's rate distortion theory [5, 3]. This paper attempts to unify the two formulations by exhibiting a clear connection between the classical rate distortion problem and the maximum likelihood mixture estimation problem using new insights based on Bregman divergences. This connection between rate distortion theory and maximum likelihood learning is significant as there exist a number of compression techniques related to the rate distortion theory such as coding using side information etc., that could potentially lead to new learning techniques or new interpretations of existing techniques.

In this paper, we analyze the rate distortion problem [3] for Bregman distortion measures [1], which include most distortion functions that are used in practice, e.g., squared Euclidean distance, KL-divergence, Itakura-Saito distance. For this case, we obtain an analytic lower bound on the rate distortion function, which is similar to the Shannon lower bound [3] for difference distortion measures. We call this lower bound, the *Shannon-Bregman lower bound* and prove that the optimal reproduction alphabet is discrete, i.e., consists of isolated singularities, unless the Shannon-Bregman lower bound is tight. Further, using the Bolzano-Weierstrass theorem, we show that for a source with bounded support, the rate distortion problem can either be analytically resolved or the optimal solution is based on a finite reproduction alphabet. Motivated by this result, we focus only on the second scenario, which is not analytically resolvable and requires numerical computation. We consider a variational problem that simultaneously addresses the problem of optimizing the rate as well as finding the optimal finite reproduction alphabet for a given distortion. This variational problem was previously analyzed in [4] for general distortion measures, but the complete algorithm and update steps were derived only for the squared error distortion measure. In this paper, we prove that a similar alternate minimization algorithm and update steps are applicable to all Bregman distortion measures.

We also provide an alternate interpretation of this rate distortion problem in terms of balancing the trade-off between compression and the loss in Bregman information [2] and show that the information bottleneck method [7] follows as an interesting special case when the sufficient statistic representation for the source entities is the conditional distribution of another random variable and the distortion measure is chosen to be the KL-divergence. Finally, we demonstrate that this variational rate distortion problem for a specified Bregman divergence is exactly equivalent to the maximum likelihood mixture learning problem based on a uniquely determined exponential family when the empirical distribution of the mixture learning problem is identical to the source distribution of the rate distortion problem.

The above equivalence may seem intuitively reasonable given that the algorithms for rate distor-

tion computation [4, 3] are similar to the expectation maximization algorithm for mixture model learning. However, these two problems arise from conceptually different goals. In the rate distortion problem, the objective is to optimize the rate distortion trade-off whereas in the mixture estimation problem the goal is to optimize the log-likelihood of the observed data. The connection between these two problems can be fully established only with the sufficient statistic representation of the corresponding random variables so that the distributions under consideration belong to the exponential family. Further, a mapping between distortions as used in rate distortion theory and the log likelihood of observations under modeling assumptions is needed - we have achieved this using a bijection between Bregman divergences and exponential family distributions [2]. It is important to note that special cases of the connection have been discovered by researchers in the recent past. For example, the connection between maximum likelihood learning of a mixture of multinomial distributions and the information bottleneck method, which was originally inspired by rate distortion theory was observed by [6].

## References

- [1] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.
- [2] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. Technical Report TR-03-19, Department of Computer Science, University of Texas at Austin, 2003.
- [3] T. Berger and J. D. Gibson. Lossy source coding. *IEEE Transactions on Information Theory*, 44(6):2691–2723, 1998.
- [4] K. Rose. A mapping approach to rate-distortion computation and analysis. *IEEE Transactions on Information Theory*, 40(6):1939–1952, 1994.
- [5] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record, Part 4*, pages 142–163, 1959.
- [6] N. Slonim and Y. Weiss. Maximum likelihood and the information bottleneck. In *NIPS 2002*, 2002.
- [7] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

**Topic:** Learning Algorithms

**Preference:** Oral/Poster