An Information-theoretic Framework for Optimization with Application to Supervised Learning [1]

David Miller, Ajit Rao, Kenneth Rose, and Allen Gersho
Center for Information Processing Research
Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106

This work develops a unified approach for hard optimization problems involving data association, i.e. the assignment of elements viewed as "data", $\{x_i\}$, to one of a set of classes, $\{C_j\}$, so as to minimize the resulting cost. The diverse problems which fit this description include data clustering, statistical classifier design to minimize probability of error, piecewise regression, structured vector quantization, as well as optimization problems in graph theory, e.g. graph partitioning. Whereas standard descent-based methods are susceptible to finding poor local optima of the cost, the suggested approach provides some potential for avoiding local optima, yet without the computational complexity of stochastic annealing.

The approach we develop is based on ideas from information theory and statistical physics, and builds on the work of Rose, Gurewitz, and Fox for clustering and related problems [1]. The optimization problem is embedded within a framework in which data are assigned to classes *in probability*, with Shannon's entropy measure used to control the level of uncertainty or randomness in the assignments. We first address "unconstrained" assignment problems such as data clustering and graph partitioning, in which the data elements are freely assigned to any class, specifiable by binary 0-1 assignment variables. We consider the joint distribution over all possible assignments, $P[x_1 \in C_{j(1)}, \ldots, x_N \in C_{j(N)}]$, and choose it to minimize the expected assignment cost $< E >$, given a constraint on Shannon's entropy, H. Thus, we seek the *best* random assignments in the sense of $< E >$ for a given H. This formulation is equivalently stated by invoking the maximum entropy principle, but the former description is more appealing for optimization. The constrained minimization is equivalent to the unconstrained minimization of the Lagrangian: $L \equiv \beta < E > -H$, where β is the Lagrange multiplier controlling $< E >$ and H. Physical inspiration for minimizing L is obtained by recognizing that it is the Helmholtz free energy of a simulated system, with $< E >$ the "energy" and $\frac{1}{\beta}$ the "temperature". Thus, a deterministic annealing approach is naturally suggested, wherein, starting from high temperature ($\beta = 0$), the cost and randomness are reduced with the temperature. At low temperature ($\beta \rightarrow \infty$) the hard cost is minimized. Our formulation unifies the deterministic annealing method for clustering with mean-field annealing methods proposed for combinatorial optimization [2]. Moreover, the derivation provides an intuitive, yet precise description of what constitutes annealing in these optimization methods. In particular, the annealing process is characterized as a reduction in the system's entropy and expected cost through the increase of a Lagrange multiplier interpreted as the inverse temperature.

While this description may provide insights into existing methods, a more significant benefit lies in its generality, and hence its potential for stimulating development of novel optimization methods tackling heretofore unaddressed assignment problems. Of prime interest are what we will call *structurally-constrained* problems, wherein the assignments are restricted to be consistent with a (parametrized) classification rule. These problems abound in pattern recognition and source coding, and include statistical classifier design, piecewise regression, and structured vector quantization. The restricted assignments may be produced by a nearest prototype rule, a decision tree, or neural network structures such as radial basis functions or multilayer perceptrons. Thus, the previous optimization framework requires substantial extension in order to enforce the structural constraint on the assignments. To do so, we introduce an additional cost C_s, which *quantifies* achievement of the structural constraint. This cost is incorporated within a generalization of the basic formulation we have described, so that the annealing process controls $< C_s >$, as well as $< E >$ and H. A second Lagrange multiplier is identified which controls $< C_s >$. This parameter is chosen to provide the optimal "level" of structural constraint consistent with $< E >$ and H at each temperature in the annealing process. At the limit $\beta \rightarrow \infty$, a "hard" classifier with the requisite structure is achieved, and the assignment cost is minimized directly. This general optimization paradigm has significant potential for outperforming descent-based approaches for structurally-constrained assignment problems. In several coming papers, these ideas are applied to the two fundamental problems of supervised learning – statistical classification and regression – as well as to the design of novel source coding structures (generalized vector quantizers), with promising results achieved in all of these domains [3], [4].

References

[1] K. Rose, E. Gurewitz, and G. C. Fox, "Vector quantization by deterministic annealing," *IEEE Trans. on Inform. Theory*, vol. 38, pp. 1249–1258, 1992.

[2] G. L. Bilbro, W. E. Snyder, and R. C. Mann, "Mean-field approximation minimizes relative entropy," *Journal of the Opt. Soc. of Amer.*, vol. 8, pp. 290–294, 1991.

[3] D. Miller, A. Rao, K. Rose, and A. Gersho, "An information-theoretic framework for optimal statistical classification." (Submitted for publication.), 1995.

[4] A. Rao, D. Miller, K. Rose, and A. Gersho, "Generalized vector quantization." (To be submitted for publication.), 1995.

[1] This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, Rockwell International Corporation, Hughes Aircraft Company, Echo Speech Corporation, Signal Technology Inc., Lockheed Missile and Space Company and Qualcomm, Inc.