M.I.T Media Laboratory Perceptual Computing Section Technical Report No. 289
Also appears in the Proceedings of the IEEE First International Conference on Image Processing, Austin, Texas, November 1994.

Exaggerated consensus in lossless image compression

Kris Popat and Rosalind W. Picard
Room E15-383, The Media Laboratory
Massachusetts Institute of Technology
20 Ames St., Cambridge, MA 02139
popat@media.mit.edu   picard@media.mit.edu

Abstract

Good probabilistic models are needed in data compression and many other applications. A good model must exploit contextual information, which requires high-order conditioning. As the number of conditioning variables increases, direct estimation of the distribution becomes exponentially more difficult. To circumvent this, we consider a means of adaptively combining several low-order conditional probability distributions into a single higher-order estimate, based on their degree of agreement. Though the technique is broadly applicable, image compression is singled out as a testing ground of its abilities. Good performance is demonstrated by experimental results.

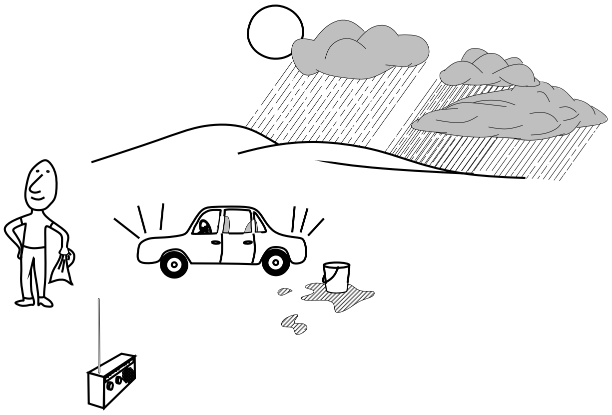
1 Introduction

In various image processing tasks, the ability to make good use of contextual information can be an important factor in determining performance. This is particularly apparent in the case of lossless image compression, where each pixel is entropy coded according to an estimated probability mass function (PMF), conditioned on contextual information that will be available at the time of decoding. Usually this information is in the form of a subset of the previously encoded pixels, or *causal neighborhood,* around the pixel to be encoded. In principle, the larger this neighborhood, the greater the compression. In practice, however, there are difficulties with large neighborhoods. The difficulties can be traced to the fact that the number of possible values (states) of the neighborhood increases exponentially with neighborhood size.

In previous papers we proposed a probability model that mitigates some of these difficulties[1, 2]. The model used clustering to summarize relevant information in the training data, and exploited the smoothness of the underlying probability law to effectively *interpolate* probability between conditioning states. When applied to lossless image compression, the technique allowed processing with large neighborhoods. However, it was found that the compression ratios achieved were only about as good as those reported for other techniques. One explanation is that, although the model allows for large neighborhoods, to adequately approximate the corresponding high-dimensional probability law would require a source-dependent degree of complexity that may be impractical.

This work was supported in part by HP Labs and NEC Corp.

In this paper we consider a different attack on the large-neighborhood problem. The next section motivates the approach through a quasi-real-world example involving strange weathermen. The idea is to obtain a large-neighborhood PMF estimate by combining several individual PMFs, each conditioned on a different small neighborhood. This problem is similar to that of integrating data from multiple sensors, or of making decisions based on accumulated evidence. These problems are ill-posed, in the sense that there are many possible "right answers" consistent with the given constraints. There will be infinitely many possible large-neighborhood PMFs consistent with a given set of small-neighborhood ones. How does one choose among them?



"... continued sunshine this afternoon, highs in the 70's."

The two principles given below can provide some guidance; in particular, they suggest the following procedure: combine the given PMFs by first taking their pointwise geometric mean, then either exaggerating or understating the shape of the result, according to whether the conditional PMFs agree or disagree with one another.

Principle 1. *If an event is impossible when conditioned on A, then it is impossible when conditioned on both A and B, regardless of B.*

This is a simple consequence of probability theory, and implies that each conditional PMF should have "veto power." This will be the case if a product rule of combination is employed, which is suggestive of the geometric

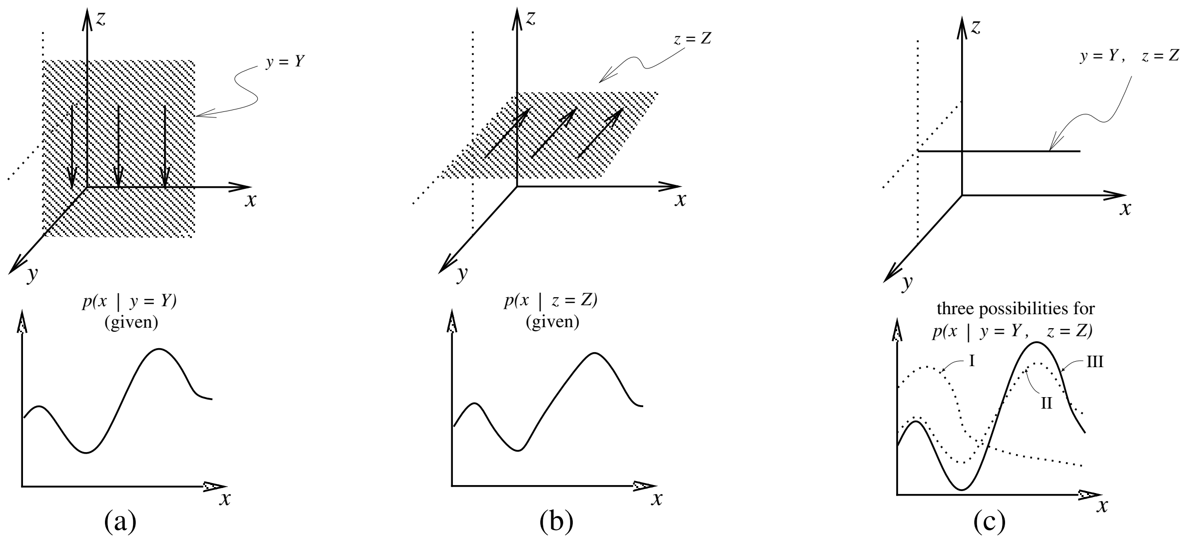


Figure 1: An example of exaggerated consensus in three dimensions. The arrows indicate the direction of integration. The lower graph in (c) shows three of the many possible shapes for $p(x|y = Y, z = Z)$, each consistent with the given $p(x|y = Y)$ and $p(x|z = Z)$. Arguably, the solid curve requires a less contrived $p(x, y, z)$ than either dotted curve.

mean or similar function.

Principle 2. *If the conditional PMFs agree (are nearly identical), then it is reasonable to assign even greater probability to those events deemed already probable in their consensus, and to assign even less probability to those events deemed already improbable.*

Unlike the first, this principle is not a consequence of probability theory, but rather it expresses a belief about what real-world joint probability distributions tend to be like. In fact, it is easy to concoct hypothetical counterexamples — joint distributions for which this principle fails entirely. However, we believe that such distributions are atypical in the applications that interest us. The next section provides a heuristic justification of this principle, while further justification is provided by the positive experimental results presented in Section 4.

2 Strange Weathermen and Least-Contrived Joint Densities

The plausibility of Principle 2 can be established by means of a somewhat fanciful example. Imagine two good but eccentric weathermen, one of whom makes her prediction of tomorrow's weather solely on the basis of today's average humidity h, ignoring all other factors. The other bases his prediction solely on today's average temperature t, while also ignoring everything else. Let R denote the event "it will rain tomorrow." Suppose that on a particular day, weatherman 1 asserts that $\Pr(R|h = H) = 0.6$, while weatherman 2 asserts that $\Pr(R|t = T) = 0.7$. Assuming for the moment that both assertions are as reliable as possible given their limited conditioning information, how might a person combine them into a single probability estimate, $\Pr(R|h = H \text{ and } t = T)$?

Since the two weathermen reach qualitatively the same conclusion — namely, that R is likely to be true — even though they base their predictions on entirely different information, it seems natural to assign an even higher probability to R than is assigned by either weatherman individually. That is, $\Pr(R|h = H \text{ and } t = T)$ should be assigned some higher value like 0.75 or 0.8, instead of the average value 0.65.

Had we not been talking about probability, but instead about some measurable physical quantity, our conclusion might have been quite different. If one lab tells you that your cholesterol level is 180 and another tells you that it's 190, you certainly would not conclude from this that it is 200. In such cases, averaging *would* make more sense.

Note the key role played here by the assumed unrelatedness of h and t. For if they had been strongly related, then the two assertions would no longer have provided independent confirmation of the likelihood of R. In the extreme case where h and t completely determine one another, simple averaging would again make more sense than exaggerating.

We now consider one more example, this one more abstract but perhaps more relevant to image processing. Let x, y, and z be random variables that obey an unknown joint density function $p(x, y, z)$. Suppose that the two conditional densities $p(x|y = Y)$ and $p(x|z = Z)$ are both known reliably for particular observations Y and Z, and that we wish to estimate $p(x|y = Y, z = Z)$ by combining the two conditionals in some way. This problem is ill-posed in the sense that there are infinitely many different possible joint densities $p(x, y, z)$, each consistent with the stated constraints (i.e., the given conditional densities). Hence, infinitely many functions $p(x|y = Y, z = Z)$ are possible.

Suppose, however, that the two given conditional densities happen to agree in shape, for instance as shown in the lower graphs in Figure 1 (a) and (b). Each corresponds to the integration of the joint density along the plane as indicated in the corresponding top graph. Again, the ill-posed question we would like to answer is, "what is the joint density along the line of intersection $(y = Y, z = Z)$, shown in the top graph in Figure 1 (c)?" Of the infinitely many possibilities, three are shown in the corresponding lower graph, representing the range of possible behavior. Curve I differs drastically in shape from the two first-order conditional densities; curve II strongly resembles the conditionals; and curve III is an exaggerated version of the conditionals. Let's consider in turn the implications of each of these being correct.

Curve I seems unlikely, since it bears no resemblance to either of the two observed first-order conditionals. For it to be correct, the joint density would have to possess a very peculiar structure. Along each of the two integration planes, $p(x, y, z)$ would have to integrate to the common shape of the given conditionals. Yet it would have to have a very different shape along the intersection line. Although certainly possible, this calls for extremely coincidental behavior of $p(x, y, z)$ in very different regions of space. Such behavior is even less plausible when the scenario is extended to higher dimensions.

Though not as obvious, curve II (the average of the given first-order conditionals) also calls for $p(x, y, z)$ to have a peculiar structure. For if no such structure were present, the effect of integrating would be to average over the many different shapes assumed by $p(x, y = Y', z = Z')$ at different values of (Y', Z'), thereby *moderating* whatever the shape happened to be at the intersection line. The given first-order conditionals, which are the results of the integration, would then be moderated versions of the function we're after, not replicas of it. Curve II requires that this moderation *not* occur, that is, it requires $p(x, y, z)$ to have a peculiar structure.

Among the three, only curve III, which exaggerates the shape of the first-order conditionals, is consistent with the expected moderating effect of integration. The exaggerated shape is consistent with a joint density that has no particular assumed structure — one less contrived than the sorts of joint densities called for by curves I and II.

"Less contrived" does not necessarily imply "smoother." Rather, a joint density which exaggerates along the intersection line is less contrived because it offers a simpler explanation for why the two first-order conditionals match each other: they inherit their common shape from the same region of probability space, namely, the region close to the intersection line. Such a joint density may or may not be smoother than others that satisfy the same constraints.

As in the weatherman example, our conclusion would have been different had y and z been strongly related (e.g., if they had been different names for the same variable), or had the shapes of the given conditionals not coincided (lack of consensus). In either of these situations, averaging would be more appropriate than exaggeration.

The words "least contrived" are suggestive of a maximum entropy formulation. Maximum entropy techniques remove unwanted degrees of freedom in underconstrained problems by finding the distribution which is closest to uniform while meeting the given constraints. The justification of maximum entropy techniques in general is a subject of debate, but our concern is only with the following question: when used in combining conditional PMFs, does it result in exaggeration? The answer is "yes," at least in some cases where exaggeration seems justified.[1] The present paper is concerned with "proof of principle" of exaggerated consensus, for which a more heuristic treatment suffices.

[1] Based on ongoing work with Ronald A. Christensen and William T. Freeman.

3 Application to lossless grayscale image compression

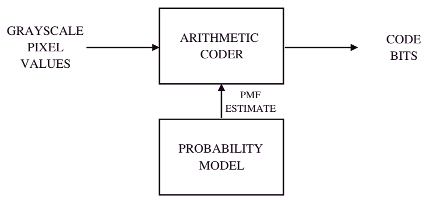


Figure 2: A lossless compression system for grayscale images.

Some applications require that an image be compressed with absolutely no distortion. Compression schemes which preserve the image exactly are termed *lossless*.

While lossy image compression systems typically compress by a factor of 10 or more while maintaining excellent image quality, lossless systems tend to give much worse compression factors (typically less than 2, see for example [3]). The exact amount of compression achieved depends on both the image's true statistical properties and the compression system's ability to exploit them. The first is beyond our control; the second is not.

Consider the lossless compression system illustrated in Figure 2. The efficiency of such a system depends on the quality of the PMF estimate provided to the arithmetic coder by the probability modeling unit for each pixel to be encoded[4].

Let x denote the next pixel to be encoded, and let $\mathcal{X}$ denote the set of values that can be assumed by x. For example, $\mathcal{X} = \{0, \ldots, 255\}$. Let $\mathcal{N}_1, \ldots, \mathcal{N}_J$ denote small causal neighborhoods of x, and let $\mathcal{N}$ denote their union. Assume that $\{\mathcal{N}_j\}$ are disjoint, so that each provides distinct contextual information about x. Since the neighborhoods are small, reliable estimates of $p(x|\mathcal{N}_j)$ are readily obtained in a variety of ways, for example by using the cluster-based probability model[2]. We therefore assume that such reliable estimates are available. We wish to estimate $p(x|\mathcal{N})$ by combining the small-neighborhood conditional PMF estimates in a manner consistent with the principles set forth in Sections 1 and 2.

To this end, we define a measure of agreement by summing the pointwise geometric mean of the conditional PMFs over x. This quantity, which we denote ρ, can be recognized as the Bhattacharyya coefficient [5]:

$$\rho = \sum_{x \in \mathcal{X}} \Big[\prod_{j=1}^{J} p(x|\mathcal{N}_j) \Big]^{1/J}.$$

It is easy to verify that $0 \leq \rho \leq 1$. When ρ is close to unity, all of the conditional PMFs agree. On the other hand, when ρ is close to zero, at least one conditional PMF is in strong disagreement with all of the others. Large values of ρ indicate consensus; small values indicate disagreement.

Motivated by our previous discussion, we define an *ex-*

aggerated consensus estimate $\tilde{p}(x|\mathcal{N})$ as

$$\tilde{p}(x|\mathcal{N}) = C\Big[\prod_{j=1}^{J} p(x|\mathcal{N}_j)\Big]^{\gamma(\rho)/J},$$

where C normalizes the estimate to make it a valid PMF, and $\gamma(\rho)$ is an exaggeration function that increases as ρ increases.

A good choice for $\gamma(\rho)$ can be obtained empirically in the following way. First, partition the range of ρ into several subintervals, in such a way that ρ is about equally likely to fall into each (as determined empirically). Next, using a test image, compute the average bit rate for each ρ-subinterval, using several candidate values of γ. For each subinterval, choose the value of γ which minimizes average bit rate. The resulting exaggeration function $\gamma(\rho)$ will be specifically tailored to that test image. If two-pass encoding is permitted, then this image-specific $\gamma(\rho)$ can be computed during the first pass and transmitted to the decoder as header information, at negligible extra cost in bit usage. Alternatively, a single, *universal* exaggeration function $\gamma(\rho)$ can be used for all images. Results for both approaches are presented.

4 Experimental results

Using the four 1-pixel conditioning neighborhoods shown in Figure 3, we obtained estimates of $p(x|\mathcal{N}_j)$, $j = 1, \ldots, 4$. Because of the low dimensionality of the probability space (each neighborhood has only one conditioning pixel), simple histogramming was selected as a reliable means of estimation, and was used in obtaining all of the results presented in this section. Empty histogram bins were avoided by initializing all bins to 1 instead of 0.

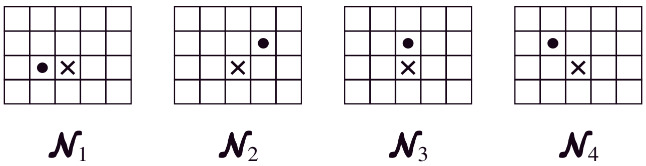


Figure 3: Conditioning neighborhoods for use in lossless image compression. In each case, the pixel marked '×' is to be encoded, and the pixel marked '•' is the conditioning value.

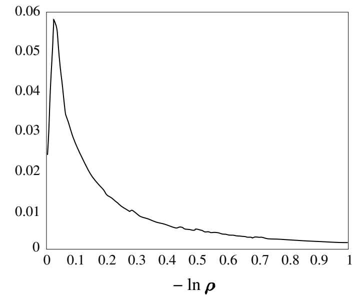


Figure 4: Normalized histogram of $-\ln\rho$, averaged over all test images.

Had the conditioning neighborhoods been larger, histogramming would have been infeasible, and a cluster-based kernel estimate (as in [1]) would be more appropriate. As a check, we repeated several of the compression experiments presented in this section using a cluster-based estimate, and found that the results were about the same.

TABLE 1: Estimated Bits per Pixel

IMAGE	0-ENT	$p(x\|\mathcal{N}_1)$	AM	GM	EC-1P	EC-2P
Al	7.70	4.89	4.74	4.73	4.42	4.41
aero2	7.30	5.04	5.09	5.10	4.74	4.73
b2	7.35	4.71	4.59	4.60	4.25	4.23
baboon	6.71	6.51	6.02	5.95	5.94	5.91
bank.512	7.66	4.98	4.68	4.74	4.33	4.32
cman	6.90	5.50	4.96	4.93	4.74	4.73
couple	7.08	4.33	4.36	4.37	3.64	3.54
crowd	7.48	6.89	6.14	6.02	6.00	5.97
einsteinB	6.87	4.83	4.77	4.78	4.27	4.24
face	7.30	5.25	4.75	4.75	4.62	4.59
fruit	6.34	5.44	4.97	4.96	4.70	4.66
girl.512	7.08	4.54	4.53	4.51	3.95	3.92
girl2k	7.53	5.52	5.19	5.18	4.85	4.83
hat	7.69	4.82	4.64	4.63	4.35	4.35
jet	5.57	4.63	4.45	4.47	4.08	4.07
kids	7.16	5.12	4.86	4.83	4.51	4.50
lenna	7.25	4.92	4.54	4.52	4.21	4.20
loco.512	5.91	5.01	4.56	4.55	4.49	4.43
london	7.30	4.31	4.49	4.54	4.07	4.03
mill	7.04	6.04	5.59	5.61	5.39	5.37
oleh	7.46	4.73	4.63	4.62	4.15	4.12
pyramid	7.33	4.66	4.67	4.71	4.06	3.93
reagan	7.32	4.76	4.52	4.51	4.13	4.12
tek-boat	7.59	6.08	5.66	5.61	5.58	5.52
tek-cute	6.96	4.68	4.71	4.73	4.15	4.12
tek-rose	7.41	6.77	6.04	5.87	5.92	5.82
vegas	7.49	4.62	4.48	4.46	4.23	4.23
wed	7.00	4.98	4.95	4.97	4.58	4.56

Different sets of estimates were obtained for each image being tested. For each image, the training set on which the estimates were based consisted of all of the *other* available images. By carefully excluding the test image from the training set, the possibility of overtraining was eliminated. This method of testing, sometimes called the "leave-one-out method," provides a relatively conservative estimate of performance.

The exact choice of ρ-partition used in determining $\gamma(\rho)$ was found to be noncritical; hence, the same partition was used for all images: $\{0, .3, .5, .65, .75, .8, .85, .9, .95, .975, 1\}$. The finer resolution for higher values of ρ is justified by the higher frequency of those values, as evidenced in Figure 4.

Experimental lossless compression results for the exaggerated consensus procedure are given in the last two columns of Table 1; the test images are shown in Figure 5. The two-pass variation (EC-2P), which uses a different exaggeration function for each image, performs slightly better than the one-pass variation (EC-1P), which uses a single universal exaggeration function for all images. Figure 6 shows the universal as well as three image-specific exaggeration functions. For comparison, the zero-order sample entropy (0-ENT) and results for the unexaggerated geometric mean (GM), the arithmetic mean (AM), and $p(x|\mathcal{N}_1)$ alone are also listed. All bit rates are estimated by assuming ideal entropy coding for the given model; previ-

Figure 5: The images used in this study. From left to right, row 1: *Al, aero2, b2, baboon, bank.512, cman, couple.* Row 2: *crowd, einsteinB, face, fruit, girl.512, girl2k, hat.* Row 3: *jet, kids, loco.512, london, mill, oleh, lenna.* Row 4: *pyramid, reagan, tek-boat, tek-cute, tek-rose, vegas, and wed.* All are 8-bit monochrome, 512 $\times$ 512, except for *cman,* which is 8-bit monochrome 256 $\times$ 256.

ous experience with arithmetic coding leads us to believe that these estimates are reliable predictors of actual performance, typically accurate to within a tenth of a bit per pixel[6].

Note that both the arithmetic and geometric means provide significant improvement over using $p(x|\mathcal{N}_1)$ alone. The geometric mean has a slight advantage over the arithmetic mean, in accordance with Principle 1. Selectively exaggerating the shape of the geometric mean results in substantial improvement, in accordance with Principle 2. The latter effect is the major finding of the present work.

5 Conclusion

A method for combining several conditional PMFs into a single PMF estimate has been presented and justified, both heuristically and by experimental findings in the application of lossless image compression. The method involves exaggerating the shape of the consensus PMF when the given conditional PMFs agree in shape.

The technique is effective in lossless image compression, and may perform similarly well in a variety of other image processing applications, such as lossy compression, restoration, segmentation, and classification. Its applicability to other domains, such as multi-sensor integration and decision-making based on accumulated evidence, is a promising area of research.

6 Acknowledgments

Our work has benefited from discussions with Ronald A. Christensen, William T. Freeman, and Martin Bichsel.

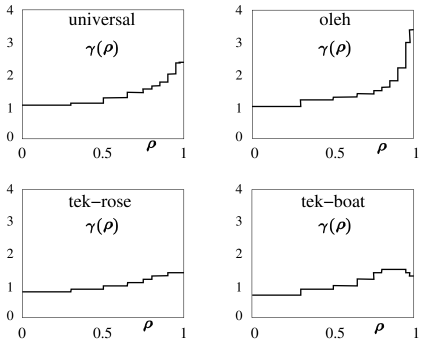


Figure 6: Universal and three image-specific exaggeration functions used in obtaining the compression results given in the last two columns of Table 1. The horizontal axis is ρ and the vertical axis is γ. The *tek-boat* function is unusual in that the degree of exaggeration diminishes slightly as ρ gets very close to 1.

References

[1] Kris Popat and Rosalind W. Picard. A novel cluster-based probability model for texture synthesis, classification, and compression. In *Proc. SPIE Visual Communications '93*, Cambridge, Mass., 1993.

[2] Kris Popat and Rosalind W. Picard. Cluster-based probability model applied to image restoration and compression. In *ICASSP-94: 1994 International Con-*

ference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, April 1994. IEEE.

[3] Glen Langdon, Amit Gulati, and Ed Seiler. On the JPEG model for lossless image compression. In *Proc. IEEE Data Comp. Conf.*, Utah, 1992.

[4] Glen G. Langdon. An introduction to arithmetic coding. *IBM J. Res. Develop.*, Mar. 1984.

[5] Thomas Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. on Commun. Technology*, COM-15(1):52–60, Feb. 1967.

[6] Kris Popat. Scalar quantization with arithmetic coding. Master's thesis, Dept. of Elec. Eng. and Comp. Science, M.I.T., Cambridge, Mass., 1990.