ON THE IMPORTANCE OF α MARGINALIZATION IN MAXIMUM ENTROPY

R. FISCHER, W. VON DER LINDEN AND V. DOSE
Max-Planck-Institut für Plasmaphysik, EURATOM Association
POB 1533, D-85740 Garching, Germany[†]

in *Maximum Entropy and Bayesian Methods*
Proceedings of the Maximum Entropy Conference 1995, Santa Fe
eds. K.M. Hanson and R.N. Silver, (1996) 229-236, Kluwer

Abstract. The correct entropic prior, computed by marginalization over the regularization parameter α, is used to invert photoemission data and to restore the famous "Susie" image. Comparison with the conventional maximum entropy procedure shows less overfitting of noise and demonstrates the residual ringing which is intrinsic to ill-posed inversion problems. An improvement to the steepest descent approximation reveals the reason for the overfitting. On top of that, the correct treatment of the regularization parameter is vital for the existence of the continuum-limit of MaxEnt.

Key words: Inverse Problem, Regularization, Entropic Prior, Image Processing

1. Introduction

The maximum entropy (MaxEnt) method is well established to recover "hidden information" from incomplete and noisy data [1]. In spite of the generality and power of MaxEnt there are still persisting problems concerning the numerical treatment. As Strauss, Wolpert and Wolf [2] pointed out, the reliability of the conventional MaxEnt procedure [3,4] depends on several conditions. In the classic MaxEnt formalism the regularization parameter α, entering the entropic prior for an N-dimensional positive additive distribution f

$$P(f|\alpha) = \frac{\exp[\alpha S(f)]}{Z_S(\alpha)}, \quad \text{where} \quad S(f) = \sum_{i=1}^{N} f_i - m_i - f_i ln(\frac{f_i}{m_i}) \quad , \tag{1}$$

has to be marginalized

$$P(f) = \int_0^\infty d\alpha P(f|\alpha)P(\alpha) \tag{2}$$

[†]Email: rrf@ipp-garching.mpg.de

since α is a nuisance parameter [5]. m is the default model and $Z_S(\alpha)$ is the normalization. According to the marginal prior (2) the posterior $P(f|D)$ does not depend on the regularization parameter

$$P(f|D) = \frac{P(D|f)P(f)}{P(D)} \tag{3}$$

In most MaxEnt-applications, however, the α-marginalization is replaced by the posterior distribution $P(f|\hat{\alpha}, D)$ where $\hat{\alpha}$ is determined by the maximum of the α-evidence $P(\alpha|D)$ [3,4]:

$$P(f|D) = \int_0^\infty d\alpha P(f|\alpha, D)P(\alpha|D) \approx P(f|\hat{\alpha}, D) = \frac{P(D|f, \hat{\alpha})P(f|\hat{\alpha})}{P(D|\hat{\alpha})} \tag{4}$$

For the evidence approximation to be valid, $P(\alpha|D)$ has to be sharply peaked in a region where $P(f|\alpha, D)$, viewed as a function of α, varies slowly. There is a further crucial approximation employed in standard MaxEnt not pointed out in [2]. Multidimensional integrals over image space f of the form

$$I = \int_0^\infty e^{\phi(f)} \frac{df^N}{\prod_i \sqrt{f_i}} \quad \approx \quad \frac{e^{\phi(f^*)}}{\prod_i \sqrt{f_i^*}} \int_{-\infty}^\infty e^{\frac{1}{2}(f-f^*)^T (\nabla\nabla^T \phi|_{f^*})(f-f^*)} df^N \tag{5}$$

are routinely approximated by 'steepest descent', i.e. by expanding ϕ to second order about its maximum f^*, extending the lower integration limits to $-\infty$, and replacing the integration measure by a constant. This approximation tends to overfit the noise for the following reason: The α-evidence $P(\alpha|D)$ is proportional to I with the exponent $\phi = \alpha S + L$ with L being the log-likelihood. Due to the divergent prefactor $1/\prod_i \sqrt{f_i^*}$ images with values close to zero are erroneously preferred, which is achieved by reducing α. The effect is amplified by the modified integration limits: the integral is overestimated by a factor 2 for each component $f_i^* \to 0$. The shortcomings of the approximate evaluation of I can be overcome by an integral-correction factor [6] which accounts for both, the integration limits and the measure.

The necessity of the explicit marginalization over α is emphasized for the first time by Bryan [5]. Strauss et al. [2] proposed a straightforward α-marginalization to find $P(f)$. In this approach the α-evidence is not required and the only remaining multidimensional integral, namely the determination of $Z_S(\alpha)$, factors into one-dimensional integrals. $P(f)$ can be tabulated efficiently since it depends only via the entropy $S(f)$ on f.

The conventional MaxEnt procedure has the further disastrous property [7] that the continuum limit does not exist. It appears that this failure occurs since α is kept fixed while changing the cell size. By marginalization over α and using Jeffrey's prior for α it can be shown that macroscopic expectation values, e.g. $\sum_{\text{domain D}} < f_i >$, exist in the continuum limit.

In the first section we apply the α-marginalization to mock data to distinguish between overfitting caused by the approximations and ringing which is intrinsic to ill-posed inversion problems. In the second section we discuss the ill-posed inversion of island-size distributions from photoemission data which depends strongly on the treatment of the regularization-parameter. Finally, we demonstrate the applicability of the α-marginalization on large data sets in the reconstruction of the "Susie" image [3].

2. Convolution data

In the upper panels of Fig. 1 a mock data set is shown consisting of a Lorentzian and a step blurred by a Gaussian. The data were distorted by Poisson noise with high (low) variance in Fig. 1a (Fig. 1c), respectively. The lower panels depict the reconstructions with the conventional MaxEnt procedure using the evidence approximation and the steepest descent procedure, the steepest descent procedure with integral corrections and the α-marginalization. On first sight the recon-

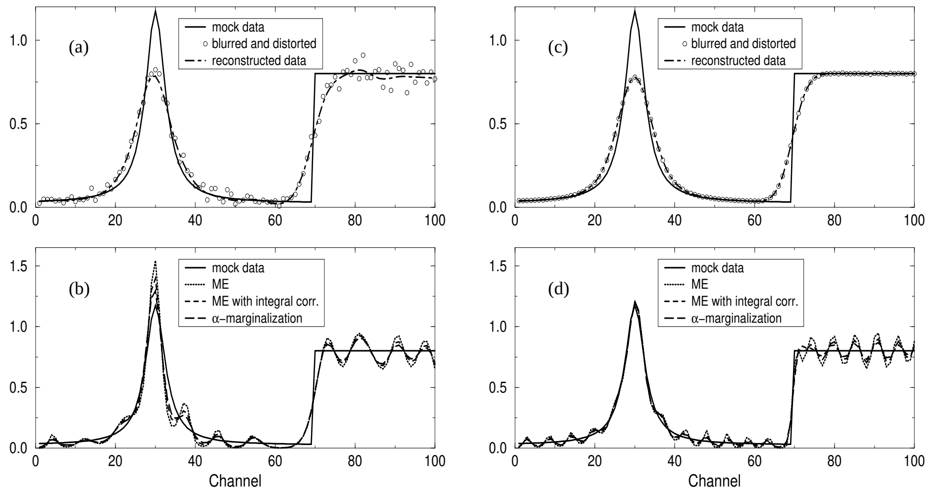

Figure 1. Upper panels: mock data (solid line) convolved with a Gaussian and distorted by Poisson noise (circles) with high (low) variance in Fig. 1a (Fig. 1c), respectively. Lower panels: MaxEnt deconvolutions with α-marginalization (long-dashed line), conventional MaxEnt procedure (dotted line) and steepest descent with integral corrections (short-dashed line, lies very close to the dotted line).

structions look frightening, the step is not reproduced properly and the image is superposed by considerable ringing.

The ringing on the plateau and in the slopes of the Lorentzian can be reduced by introducing a preblur on a "hidden image" as proposed by Skilling [8] (not shown here). But our calculations showed that the preblur has the property to set as many variables of the hidden image to very small values. It appears that this tendency is a consequence of the steepest descent approximation. Small values are erroneously preferred due to the divergent prefactor and the modified integration limits. Furthermore, sharp structures like the step cannot be reproduced with a constant preblur necessary for reducing the high ringing level. For the future, the introduction of a preblur with variable width over the image following the pixon approach [9] may be promising.

An integral-correction factor to the steepest descent approximation demonstrates the essential contribution to the overfitting. The reconstruction obtained by the integral-correction factor (short-dashed line) hardly deviates from that obtained by the α-marginalization (long-dashed line).

The reason for the disappointing ringing which even resists the α-marginalization is intrinsic to the problem. To show this analytically the problem is simplified. A Gaussian likelihood distribution $P(g|f)$ is chosen with a linear model Af and a quadratic prior distribution $P(f|\alpha, m)$

$$P(g|f) \quad \propto \quad \exp\left\{-\frac{(g-Af)^2}{2\sigma^2}\right\} \tag{6}$$

$$P(f|\alpha, m) \quad \propto \quad \exp\left\{-\frac{\alpha}{2}(f-m)^2\right\} \quad . \tag{7}$$

The posterior mode is then given by

$$-A^T g + A^T A f + \lambda(f-m) = 0 \quad \text{with} \quad \lambda = \alpha\sigma^2 \quad . \tag{8}$$

Splitting the data $g = g^x + \Delta g$ and the reconstruction $f = f^x + \Delta f$ into exact values and noise, and denoting $m = f^x + \Delta m$, the equation for the deviations reads:

$$(A^T A + \lambda)\Delta f = A^T \Delta g + \lambda \Delta m \tag{9}$$

The first term on the rhs describes the influence of the noise on the reconstruction f. For example, without regularization ($\lambda = 0$) small variations in the data result in huge variations in f if the eigenvalues of $A^T A$ fall off very fast. The ringing in Fig. 1b is mainly due to the large noise level in the data (Fig. 1a). The second term, denoting the difference between the default model m and the exact solution f^x, causes ringing even for $\Delta g \to 0$. The ringing caused by the Δm-term is shown in Fig. 1d where the noise level is very low (Fig. 1c). The ringing occurs for small λ in this limit and for $\Delta g = 0$ equation (9) yields

$$\Delta f_i \approx \sum_{\nu, \varepsilon_\nu < \lambda} \phi_i^{(\nu)} \left(\sum_j \phi_j^{(\nu)} \Delta m_j \right) \quad , \tag{10}$$

with $\phi^{(\nu)}$ being the ν-th eigenvector of $A^T A$ and ε_ν the corresponding eigenvalue. This expression predicts considerable ringing if $A^T A$ has eigenvalues smaller than λ and if the overlap between the corresponding eigenvectors and Δm, the discrepancy between default model and exact solution, is large. This is the case if the apparatus function (described by A) is much broader than the structures in Δm, i.e. structures in the image which are not accounted for by the default model. Ringing is therefore mainly due to the ill-posed nature of the inversion problem and could only be overcome by an improved default model m. Yet to a small extend, it is also influenced by the treatment of the regularization parameter (λ) as can be seen in Fig. 1c,d.

3. Island-size distributions from photoemission data

Our actual problem was the reconstruction of island-size distributions from photoemission data of ultra-thin Ag coverages on a Pd(111) surface. Without going

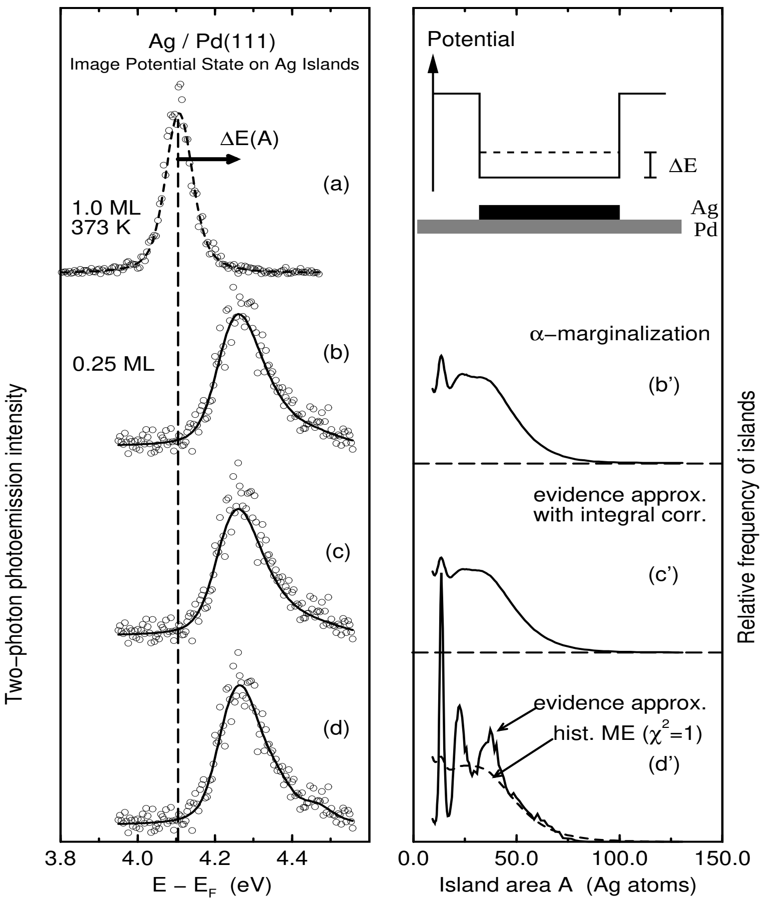

Figure 2. Photoelectron spectra of a surface state on the homogeneous Ag monolayer (a) and on one atomic layer high Ag islands at a Ag coverage of 0.25 monolayer (b)-(c) (left panel). The surface state on islands is shifted in energy relative to the state of the homogeneous Ag layer due to confinement on Ag-islands (right-panel inset). Island-size distributions derived from the experimental data are shown in the right panel for different approaches: the established evidence approximation (d), historic MaxEnt where α is determined via $\chi^2 = 1$ (dashed), the evidence approximation with integral corrections (c), and the α-marginalization (b).

into details a short summary should highlight the problem. Small amounts of Ag adsorbed on a Pd single crystal surface form islands with sizes of a few nm. On the homogeneous Ag surface an electronic surface state exists with a well defined energy and width (Fig. 2a). On small Ag islands the surface state is shifted to higher energy due to the localization of the state on these islands (Fig. 2b-d). Like a quantum mechanical particle in a potential box, here the electron of the surface state on the Ag island, the energy of the state increases with decreasing box size (Fig. 2 right-panel inset). Approximately, the energy shift is inversely proportional to the island area. This surface state is a very sensitive tool to study the surface morphology of the sample [10,11]. The energy and width of the surface state gives us information about nucleation and growth of the Ag adsorbate. In a photoemission experiment the surface state is measured with the energy distribution curves of the emitted photoelectrons.

The surface state on small Ag islands [(b)-(d)] is energetically shifted and asymmetrically broadened. The shift of the surface state is assigned to the lateral confinement and the broadening is due to the existence of a distribution of different island sizes. Hence, the energy distribution curve bears the information about the size distribution of the Ag islands.

To reconstruct the island-size distribution we model the photoemission spectrum by an integral over the energy shifted spectrum $K(E,A)$ of the homogeneous Ag film (Fig. 2a) times the island-size distribution function $f(A)$

$$G(E) = \int_0^\infty dA K(E,A) f(A) \quad . \tag{11}$$

The energy shift is calculated for a finite two-dimensional circular potential well. Further details can be found in [10–12]. The island-size distributions derived from the energy distribution curve of the experimental data are shown in the right panel of Fig. 2. Four different approaches were employed to invert eq. (11): conventional MaxEnt (d'), historic MaxEnt where α is determined via $\chi^2 = 1$, MaxEnt plus integral-correction (c'), and α-marginalization (b'). The reconstruction in data-space to the various approaches are shown as solid lines in the left panel. The conventional MaxEnt approach (d') clearly yields an overfitted solution. The exact approach (b') is much smoother. The integral corrected version of the conventional MaxEnt (c') is close to the exact result. Even historic MaxEnt yields a much better result than the approximate classic MaxEnt.

The exact approach allows to determine an effective α_P, although it marginalizes over α [2]. Relative to α_P the regularization parameter is for (d') a factor 17 smaller and for (c') a factor 1.8 larger. The surprising fact was that if we take the effective α_P of the α-marginalization as the optimum $\hat{\alpha}$ in the conventional MaxEnt approach we get the result of the α-marginalization. From this we conclude that the α-evidence $P(\alpha|D)$ is indeed sharply peaked and all the problems are caused by the approximate treatment of the evidence-integrals. The conditions for the evidence analysis are, at least in this case, fulfilled. The small remaining discrepancy between MaxEnt plus integral correction and the exact approach is supposedly due to the Taylor-expansion of the exponent Φ.

The dominant peak at an island size of 14 Ag atoms is due to the small structure in the data at an energy of 4.47 eV. The result obtained by the α-marginalization does not emphasize further island sizes. The existence of a sub-structure on a broad distribution was surprising and is interpreted as a favored magic island size. The overfitted solution of the conventional MaxEnt approach yields two more magic island sizes due to the approximations. This example shows the necessity for the correct regularization with the α-marginalization.

4. Image reconstruction with "Susie"

To assess the α-marginalization we restored the famous "Susie"-image which is digitized on a 128×128 pixel grid with grey-level values between 40 and 255 (Fig. 3a). This is a traditional example for MaxEnt image processing [3, and references therein]. For comparison with previous approaches we blurred the picture by a 6-pixel radius Gaussian point-spread function and added noise of unit variance (Fig. 3b). The reconstruction obtained by the conventional MaxEnt approach shows nasty overfitting (Fig. 3d). An additional numerical problem of the conventional MaxEnt procedure is the evaluation of the logarithm of the determinant of the 16384×16384 Hessian matrix which is numerically rather cumbersome. This task can be tackled either by Monte Carlo techniques or it can be avoided [3] by using a quadratic prior. A great advantage of the α-marginalization is that this difficulty does not arise in the first place.

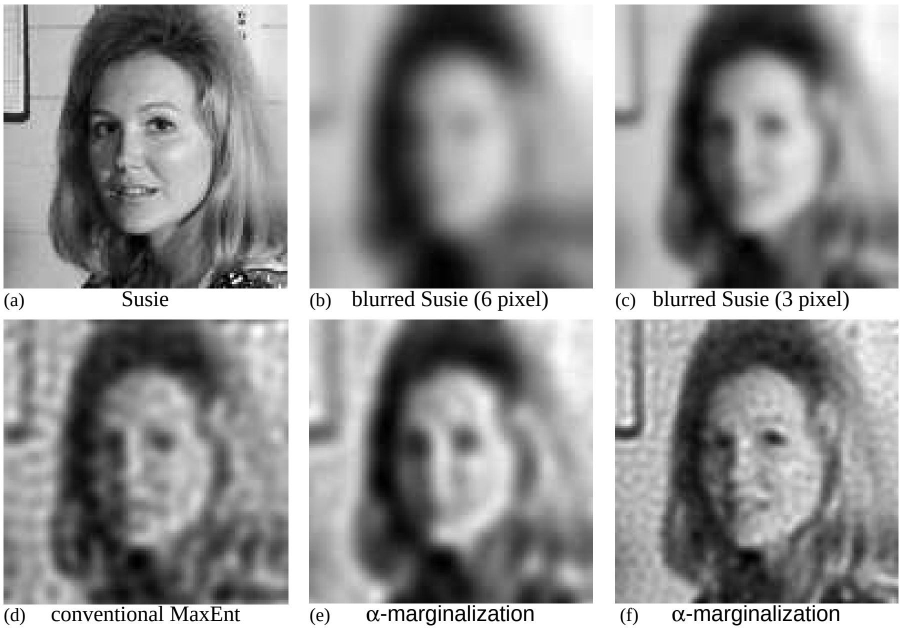

Figure 3. Upper panel: image of "Susie" (a), blurred by a 6[3]-pixel Gaussian PSF (b) [(c)], respectively. Lower panel: MaxEnt reconstructions using the evidence approximation (d) and the α-marginalization [(e), (f)].

The result obtained by the α-marginalization (Fig. 3e) is much more pleasant due to the reduced ringing level. As Strauss et al. supposed the reconstruction of Susie based on the evidence approximation is overfitted. Nevertheless, the inversion-intrinsic ringing is still present. The intrinsic ringing increases if the structures get sharper. To show this we performed a 3-pixel blurred image of Susie (Fig. 3c). The optimal reconstruction (Fig. 3f) obtained by the α-marginalization is still superposed by considerable ringing. This can only be avoided by a more elaborate default model.

5. Summary

We have demonstrated that a great deal of overfitting, generally observed in MaxEnt images, is due to the approximate evaluation of the α-evidence and the replacement of the α-marginalization by α-evidence analysis. It appears that the overfitting of noise can be eliminated by the α-marginalization while the ringing, caused by the ill-posed inversion problem, is retained. The correct treatment of the regularization parameter has also a strong impact on adjusting other hyperparameters, e.g. the width of the preblur proposed by Skilling. It is also vital for the correct continuum-limit of the theory. For the reconstruction of island-size distributions from photoemission data the exact regularization turned out to be essential for the physical problem under consideration. The applicability of the α-marginalization to large data sets is demonstrated by the reconstruction of the Susie image.

References

1. S. F. Gull and J. Skilling, "Maximum entropy method in image processing," *IEE Proc.*, **131 F**, pp. 646–659, 1984.
2. C. E. M. Strauss, D. H. Wolpert and D. R. Wolf, "Alpha, evidence, and the entropic prior," in *Maximum Entropy and Bayesian Methods (1993)*, G. Heidbreder, ed., Kluwer Academic, Dordrecht, 1995.
3. S. F. Gull, "Developments in maximum entropy data analysis," in *Maximum Entropy and Bayesian Methods*, J. Skilling, ed., pp. 53–71, Kluwer Academic, Dordrecht, 1989.
4. J. Skilling, "Quantified maximum entropy," in *Maximum Entropy and Bayesian Methods*, P. F. Fougère, ed., pp. 341–350, Kluwer Academic, Dordrecht, 1990.
5. R. Bryan, "Maximum entropy analysis of oversampled data problems," *Eur. Biophys. J.*, **18**, pp. 165–174, 1990.
6. W. von der Linden, R. Fischer, and V. Dose, "Evidence integrals," *In these proceedings*, 1995.
7. J. Skilling and S. Sibisi, "Prior distributions on measure space," *In these proceedings*, 1995.
8. J. Skilling, "Fundamentals of MaxEnt in data analysis," in *Maximum Entropy in Action*, B. Buck and V. A. Macaulay, eds., pp. 19–40, Clarendon Press, Oxford, 1991.
9. R. Puetter, "Pixon-based image reconstruction in high energy astrophysics," *In these proceedings*, 1995.
10. R. Fischer, S. Schuppler, N. Fischer, Th. Fauster and W. Steinmann, "Image states and local work function for Ag/Pd(111)," *Phys. Rev. Lett.*, **70**, pp. 654–657, 1993.
11. R. Fischer, Th. Fauster and W. Steinmann, "Threedimensional localization of electrons on Ag islands," *Phys. Rev. B*, **48**, pp. 15496–15499, 1993.
12. R. Fischer, "submitted to Surf. Sci."