QUANTITATIVE L^2 APPROXIMATION ERROR OF A PROBABILITY DENSITY ESTIMATE GIVEN BY IT SAMPLES

Thierry Blu and Michael Unser

Biomedical Imaging Group
Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland
e-mail:thierry.blu@epfl.ch,michael.unser@epfl.ch

ABSTRACT

We present a new result characterized by an exact integral expression for the approximation error between a probability density and an integer shift invariant estimate obtained from its samples. Unlike the Parzen window estimate, this estimate avoids recomputing the complete probability density for each new sample: only a few coefficients are required making it practical for real-time applications.

We also show how to obtain the exact asymptotic behavior of the approximation error when the number of samples increases and provide the trade-off between the number of samples and the sampling step size.

1. INTRODUCTION

Probability density estimation is a key issue in many applications such as pattern recognition, denoising, classification, or multimodal image/volume registration [1]. The problem amounts to finding a good estimate of $p(x)$ given a set $\{x_i\}_{i=1,\ldots,N}$ of independent realizations—samples—of the random variable x. A standard non-parametric density estimate is the point probability density

$$p_\delta(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i) \qquad (1)$$

from which it is customary to build the kernel estimator [2] $p_{\text{est}}(x) = h^{-1}\chi(h^{-1}x) * p_\delta(x)$. In this expression, the Parzen window $h^{-1}\chi(h^{-1}x)$ is a positive function with normalized integral, and the parameter h controls its size.

In this paper, we study another histogram-like estimate that is closely related to wavelet probability density estimations [3, 4, 5, 6, 7], and we evaluate the probabilistic expectation of the $\mathbf{L}^2$-approximation error between the true density and this estimate. We obtain an exact expression which allows us to compute the optimal bin size given the number of available samples and we give some exact asymptotic results that are similar to [5, 6].

2. DESCRIPTION

Instead of filtering the point distribution (1) by a Parzen window, we perform a shift-invariant linear approximation—e.g., a projection—of p_δ onto some function space $\mathcal{V}_h = \text{span}_{n\in\mathbb{Z}}\{\varphi_{h,n}(x)\}$ according to the formula

$$p_{\text{est}}(x) = \mathcal{Q}_h p_\delta(x) = \sum_{n\in\mathbb{Z}}\langle p_\delta, \tilde{\varphi}_{h,n}\rangle\varphi_{h,n}(x) \qquad (2)$$

where we use the notation $f_{h,n}(x) = h^{-1/2}f(h^{-1}x-n)$. Here, $\tilde{\varphi}$ and φ are functions which may or may not have good approximation properties—e.g., approximation order, biorthonormality [8, 9]. Note that the standard histogram is obtained by choosing $\tilde{\varphi} = \varphi = \text{rect}$.

The expression (2) is particularly interesting from an algorithmic point of view because, unlike the Parzen-window expression, it is not necessary to recompute the estimator every time a new sample is received, but only to update the coefficient $c_n^{(N)} = \langle p_\delta, \tilde{\varphi}_{h,n}\rangle$ of the expansion (2) according to the formula

$$c_n^{(N+1)} = \frac{N}{N+1}c_n^{(N)} + \frac{1}{N+1}\tilde{\varphi}_{h,n}(x_{N+1}).$$

Note that there are only a finite number of coefficients to update if $\tilde{\varphi}$ has a finite support, which is what we will choose in practice. This algorithm is particularly well-suited for real-time processing.

Another interesting property of this estimator is that it preserves the discrete moments of the probability density up to degree $L-1$

$$\int x^l p_{\text{est}}(x)\,\mathrm{d}x = \frac{1}{N}\sum_{n=1}^{N}x_i^l$$

whenever $\tilde{\varphi}$ has approximation order L and $(\varphi,\tilde{\varphi})$ is biorthonormal—i.e., $\hat{\tilde{\varphi}}(\omega) = \delta_n + O((\omega-2n\pi)^L)$ and the cross-correlation $\langle\tilde{\varphi}_{h,n},\varphi_{h,n'}\rangle = \delta_{n-n'}$ for all integers n, n'. In particular, it is interesting to choose φ and $\tilde{\varphi}$ such that they span the same space and are biorthonormal: in this configuration, the operator $\mathcal{Q}_h$ is an orthogonal projection onto

this space. Moreover, since a probability density should always sum up to 1, we will always impose that $\tilde{\varphi}$ is at least of approximation order 1.

3. MAIN RESULTS

Following [8, 9], we observe that if we shift an $\mathbf{L}^2$ function f by τ then the approximation error $\|f_\tau - \mathcal{Q}_h f_\tau\|_{\mathbf{L}^2}$ is h-periodic. Here, we have denoted by f_τ the shifted function $f(\cdot - \tau)$. This incites to define the following *shift-averaged* approximation-error measure

$$\eta(f)^2 = \frac{1}{h}\int_0^h \|f_\tau - \mathcal{Q}_h f_\tau\|_{\mathbf{L}^2}^2 \,\mathrm{d}\tau.$$

We have shown in [9] that, under slight conditions on f (Sobolev smoothness $> 1/2$), on $\tilde{\varphi}$ (bounded Fourier transform) and on φ (Riesz basis condition), we have

$$\eta(f)^2 = \frac{1}{2\pi}\int |\hat{f}(\omega)|^2 E(h\omega)\,\mathrm{d}\omega \qquad (3)$$

where we have defined the Fourier approximation kernel

$$E(\omega) = |1 - \hat{\tilde{\varphi}}(\omega)^* \hat{\varphi}(\omega)|^2 + |\hat{\tilde{\varphi}}(\omega)|^2 \sum_{n\neq 0} |\hat{\varphi}(\omega + 2n\pi)|^2.$$

The shift-averaged approximation error $\eta(f)$ is very close (and even equal) to the true $\mathbf{L}^2$ approximation error, the difference between them being of the order of h^s where s is the Sobolev regularity of the function f. Note that the quality of the approximation as $h \to 0$ is all the better as the Fourier approximation kernel cancels with a higher power at $\omega = 0$.

Because we consider a random estimate p_{est}, it is the evaluation of $\mathcal{E}\{\|p - \mathcal{Q}_h p_\delta\|_{\mathbf{L}^2}^2\}$ with which we are concerned in this paper, where $\mathcal{E}\{\cdot\}$ denotes the probabilistic expection of a random variable. Similarly as above, we observe that changing the origin of x by a multiple of h does not change this expectation. This is why we will evaluate instead the shift-averaged expression

$$\bar{\eta}(f)^2 = \frac{1}{h}\int_0^h \mathcal{E}\{\|p_\tau - \mathcal{Q}_h (p_\delta)_\tau\|_{\mathbf{L}^2}^2\}\,\mathrm{d}\tau. \qquad (4)$$

Theorem 1 *The expected approximation error of the probability density $p(x)$ using the estimate* (2) *is given by*

$$\bar{\eta}(p)^2 = \int |\hat{p}(\omega)|^2 E(h\omega)\frac{\mathrm{d}\omega}{2\pi} + \frac{1}{N}\int (1 - |\hat{p}(\omega)|^2) S(h\omega)\frac{\mathrm{d}\omega}{2\pi} \qquad (5)$$

where

$$S(\omega) = |\hat{\tilde{\varphi}}(\omega)|^2 \sum_{n\in\mathbb{Z}} |\hat{\varphi}(\omega + 2n\pi)|^2.$$

Proof: We first observe that $\mathcal{E}\{p(x) - p_\delta(x)\} = 0$ which means that p_δ is an unbiased estimate of p. This implies that $\mathcal{E}\{\|p_\tau - \mathcal{Q}_h (p_\delta)_\tau\|_{\mathbf{L}^2}^2\} = \|p_\tau - \mathcal{Q}_h p_\tau\|_{\mathbf{L}^2}^2 + \mathcal{E}\{\|\mathcal{Q}_h(p_\tau - (p_\delta)_\tau)\|_{\mathbf{L}^2}^2\}$. This expression has to be further integrated over τ according to (4).

The first term is exactly the first rhs term of (5) thanks to (3).

The second term, $\mathcal{E}\{\|\mathcal{Q}_h(p_\tau - (p_\delta)_\tau)\|_{\mathbf{L}^2}^2\}$, can be evaluated as follows. It was shown in [8] that

$$h^{-1}\int_0^h \|\mathcal{Q}_h f_\tau\|_{\mathbf{L}^2}^2\,\mathrm{d}\tau = \int |\hat{f}(\omega)|^2 S(h\omega)\frac{\mathrm{d}\omega}{2\pi}.$$

We apply this result to $f = p - p_\delta$ and take the expectation of the result. This involves the computation of $\mathcal{E}\{|\hat{p}(\omega) - \hat{p}_\delta(\omega)|^2\} = \mathcal{E}\{|\hat{p}_\delta(\omega)|^2\} - |\hat{p}(\omega)|^2$ and we have

$$\mathcal{E}\{|\hat{p}_\delta(\omega)|^2\} = \frac{1}{N^2}\sum_{i,i'=1}^{N} \mathcal{E}\left\{e^{-j\omega(x_i - x_{i'})}\right\} = \left(1 - \frac{1}{N}\right)|\hat{p}(\omega)|^2 + \frac{1}{N}$$

which finally provides the second rhs term of (5). □

We exemplify in Fig. 1 the relation between the true apoproximation error and the quantity $\bar{\eta}(p)$.

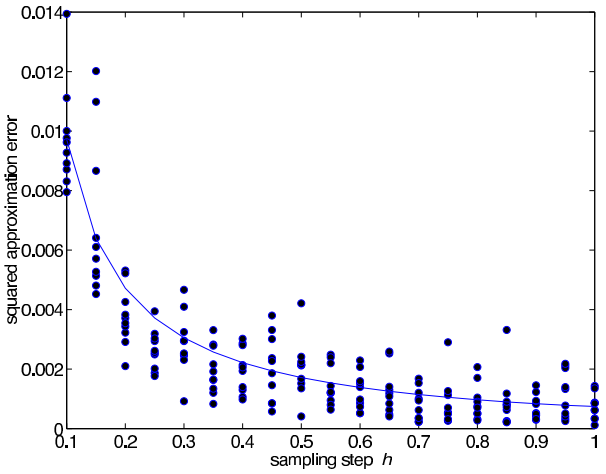

Fig. 1. Case of $N = 1000$ zero-average Gaussian data with variance 1. Comparison between the true approximation error using cubic splines and sampling step $0.1 \leq h \leq 1$ (circles) and the prediction using $\bar{\eta}(p)$.

This result is very interesting because it provides some good insight into the mixed effect of the quality of the approximation method and of the number of samples available. For instance, if we let h tend to zero, the quality of the approximation improves—smaller sampling step—which is seen on the first rhs term of (5). On the contrary, when $h \to 0$, the second rhs term tends to infinity, reflecting

the fact that $p_{\text{est}} \to p_\delta$ which is not square-integrable. Of course, if we increase N while decreasing h accordingly, it will be possible to counterbalance this effect.

Theorem 2 *Assume that the couple of functions $(\tilde{\varphi}, \varphi)$ is biorthonormal and that φ is of approximation order L. Then, for small values of h, the expected approximation error of the estimate* (2) *can be expressed as*

$$\bar{\eta}(p)^2 = C_\varphi^2 h^{2L} \|p^{(L)}\|_{\mathbf{L}^2}^2 + \frac{D_{\varphi,\tilde{\varphi}}^2}{Nh} \tag{6}$$

where $C_\varphi = \sqrt{\sum_{n\neq 0} |\hat{\varphi}^{(L)}(2n\pi)|^2}$ is the asymptotic constant for the orthogonal projection [10] and where

$$D_{\varphi,\tilde{\varphi}}^2 = \sum_{n\in\mathbb{Z}} \langle \varphi, \varphi_n \rangle \langle \tilde{\varphi}, \tilde{\varphi}_n \rangle.$$

In particular, when $(\tilde{\varphi}, \varphi)$ is biorthonormal, $D_{\varphi,\tilde{\varphi}} \geq 1$ with equality only when $\mathcal{Q}_h$ is the orthogonal projection onto $\text{span}_{n\in\mathbb{Z}}\{\varphi_n\}$.

Proof: As regards the first rhs term of (5), we already know from [9] that the asymptotic approximation error takes the form $C_\varphi h^L \|p^{(L)}\|_{\mathbf{L}^2}$. The second term can be rewritten as $h^{-1} \int S(\omega) \frac{\mathrm{d}\omega}{2\pi} - \int S(h\omega) |\hat{p}(\omega)|^2 \frac{\mathrm{d}\omega}{2\pi}$. When $h \to 0$, it is the h^{-1}-term which is dominant and we have that

$$\begin{aligned} D_{\varphi,\tilde{\varphi}}^2 &= \int S(\omega) \frac{\mathrm{d}\omega}{2\pi} \\ &= \frac{1}{2\pi} \int_0^{2\pi} \Big(\sum_{n\in\mathbb{Z}} |\hat{\tilde{\varphi}}(\omega + 2n\pi)|^2 \Big) \\ &\qquad \times \Big(\sum_{n\in\mathbb{Z}} |\hat{\varphi}(\omega + 2n\pi)|^2 \Big) \, \mathrm{d}\omega \end{aligned}$$

from which the expression for $D_{\varphi,\tilde{\varphi}}$ follows, after noticing that $\sum_{n\in\mathbb{Z}} |\hat{f}(\omega + 2n\pi)|^2$ is the discrete-time Fourier transform of the autocorrelation sequence $\{\langle f, f_n \rangle\}_{n\in\mathbb{Z}}$.

When $(\tilde{\varphi}, \varphi)$ is biorthonormal, $\sum_{n\in\mathbb{Z}} \hat{\tilde{\varphi}}(\omega+2n\pi)^* \hat{\varphi}(\omega + 2n\pi) = 1$ holds true. Using Cauchy-Schwarz inequality, we find that $D_{\varphi,\tilde{\varphi}} = 1$ iff $\tilde{\varphi}$ lives in $\text{span}_{n\in\mathbb{Z}}\{\varphi_n\}$. □

This theorem gives us the exact—asymptotic—tradeoff between the bin size (or sampling step) h and the number of samples N. We can then find the optimal couple (N, h) that will yield the smallest asymptotic error.

Corollary 1 *Within the framework of Thm. 2, the optimal couple (N, h) is tied by the relation*

$$h = \Big(\frac{N_0}{N} \Big)^{\frac{1}{2L+1}} \tag{7}$$

where

$$N_0 = \frac{D_{\varphi,\tilde{\varphi}}^2}{2L\, C_\varphi^2 \, \|p^{(L)}\|_{\mathbf{L}^2}^2}.$$

When this relation is satisfied, the approximation error $\bar{\eta}(p)$ decreases to zero according to

$$\bar{\eta}(p) = \sqrt{2L+1} \left(C_\varphi D^{2L} \|p^{(L)}\|_{\mathbf{L}^2} \right)^{\frac{1}{2L+1}} \times N^{-\frac{L}{2L+1}}. \tag{8}$$

Proof: By differentiation of (6). □

The decrease rate $N^{-L/(2L+1)}$ is of course known in the literature [4, 7] in the more general case of probability densities $p(x)$ that belong to Besov spaces. Here, we also provide exact asymptotic constants. In particular, when φ is a spline of order L—i.e., of degree $(L-1)$—and $\mathcal{Q}_h$ is the orthogonal projection onto the spline space, we have that $C_\varphi = (2\pi)^{-L} \sqrt{2\zeta(2L)}$ where $\zeta(s)$ is Riemann's zeta function $\sum_{n\geq 1} n^{-s}$. Now, for L large enough we have that

$$\bar{\eta}(p) \approx \frac{\|p^{(L)}\|_{\mathbf{L}^2}^{\frac{1}{2L+1}}}{\sqrt{2\pi}} N^{-\frac{L}{2L+1}}.$$

If instead of choosing the spline of order L, we choose the OMOMS—the shortest functions of order L having the smallest asymptotic constant—of same order [11], the asymptotic constant is now $C_\varphi = \frac{L!}{(2L+1)!\sqrt{2L+1}}$, and, for large L, we have that

$$\bar{\eta}(p) \approx \frac{\sqrt{e}\, \|p^{(L)}\|_{\mathbf{L}^2}^{\frac{1}{2L+1}}}{2\sqrt{L}} N^{-\frac{L}{2L+1}}$$

which has a much smaller constant and thus a better approximation than the spline approximation.

4. CONCLUSION

We have presented a predictive expression for the approximation error between a sampling density estimate and the true probability density. The accuracy of this expression has allowed us to quantify the respective influence of the approximation space, and the influence of a larger number of samples. In particular, we were able to obtain exact asymptotic expressions.

5. REFERENCES

[1] P. Thévenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Transactions on Image Processing*, vol. 9, no. 12, pp. 2083–2099, December 2000.

[2] E. Parzen, "On the estimation of a probability density function and the mode," *Ann. Math. Statistics*, vol. 33, pp. 1065–1076, 1962.

[3] P. Doukhan and J.R. Léon, "Déviation quadratique d'estimateurs de densité par projections orthogonales," *C. R. Acad. Sci. Série I*, vol. 310, pp. 425–430, 1990.

[4] G. Kerkyacharian and D. Picard, "Density estimation in Besov spaces," *Stat. Prob. Lett.*, vol. 13, pp. 15–24, 1992.

[5] E. Masry, "Probability density estimation from dependent observations using wavelet orthonormal bases," *Stat. Prob. Lett.*, vol. 21, pp. 181–194, 1994.

[6] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," *Ann. Stat.*, vol. 24, pp. 508–539, 1996.

[7] O. Renaud, *Density estimation with wavelets: Variability, invariance and discriminant power*, Ph.D. thesis, École Polytechnique Fédérale, Lausanne, Switzerland, 1999.

[8] T. Blu and M. Unser, "Approximation error for quasi-interpolators and (multi-) wavelet expansions," *Appl. Comput. Harmon. Anal.*, vol. 6, no. 2, pp. 219–251, March 1999.

[9] T. Blu and M. Unser, "Quantitative Fourier analysis of approximation techniques: Part I—Interpolators and projectors," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2783–2795, October 1999.

[10] M. Unser, "Approximation power of biorthogonal wavelet expansions," *IEEE Trans. Signal Process.*, vol. 44, no. 3, pp. 519–527, March 1996.

[11] T. Blu, P. Thévenaz, and M. Unser, "MOMS: Maximal-Order Interpolation of Minimal Support," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1069–1080, July 2001.