Appendix

1 Polynomial Integration using Green's Theorem

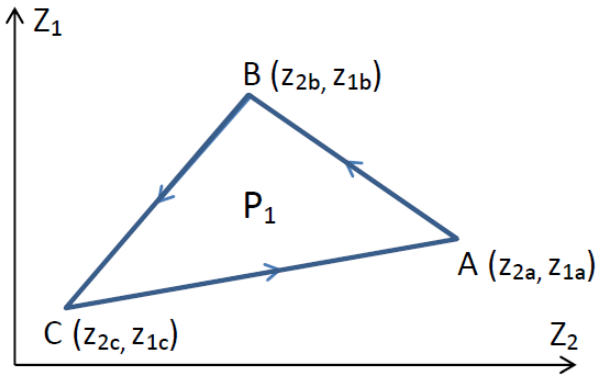

Figure 1: Image depicts the problem setting for polynomial integration over a closed polygon using the Green's theorem. P_1 is a polynomial defined over the polygon ABC. The vertex coordinates are indicated in the round brackets. Using the Green's theorem, P_1 can be integrated over the polygon ABC by evaluating the line integrals along the edges AB, BC, and CA in the counterclockwise direction (indicated by arrows) and summing them up.

Let P_1 be the polynomial function of (Z_1, Z_2) defined over the polygon ABC. In order to evaluate the integration of P_1 over the polygon ABC, we first choose the polynomials L and M such that $L = (\frac{-1}{2}) \int P_1 \, \mathrm{d}Z_1$ and $M = \frac{1}{2} \int P_1 \, \mathrm{d}Z_2$. Therefore, we have $\frac{\partial M}{\partial Z_2} - \frac{\partial L}{\partial Z_1} = P_1$. The line integral over the edge AB for the polynomial P_1 can be computed using the formula $\int_{z_{2a}}^{z_{2b}} L \, \mathrm{d}Z_2 + \int_{z_{1a}}^{z_{1b}} M \, \mathrm{d}Z_1$. By Green's theorem, the integration of P_1 over the polygon ABC can be computed as the sum of the line integrals evaluated in a counterclockwise direction along the edges AB, BC, and CA.

2 Bandwidth Estimation and Choice of Kernel

We briefly explain the process of kernel density estimation for our experiments. The choice of the kernel and the bandwidth selection are the two important parameters that affect estimation of the underlying density with nonparametric models. The bandwidth of the kernel strongly influences the estimated density function. Small bandwidth can undersmooth the estimated density, whereas large bandwidth can oversmooth the estimated density function, thus losing the important details, e.g., mode. The proposed approach for computing the ratio distribution is sensitive to the bandwidth selection. The area of each of the parallelograms shown in Fig. 2 in the paper is determined by the bandwidth of the kernel. The areas of the parallelogram falling in the domain of the ratio distribution, [0,1], determine the shape of the ratio distribution as explained in Subsubsection 3.2.1 in the paper. Thus, a poor choice of bandwidth can affect the shape of the analytically computed ratio density. Bandwidth selection is therefore an important research problem for reliable kernel density estimation.

Jones et al. [1] provide a brief overview of the various bandwidth estimation techniques. In our work, we use the plugin method proposed by Sheather and Jones [5] and the rule of thumb approach [1] for the bandwidth selection. Most of the automatic bandwidth selection methods penalize high mean integrated squared error (MISE) [4] between the estimated density and the underlying true density. The MISE in case of kernel density estimation is given by:

$$MISE(\hat{f}) = E\int(\hat{f}-f)^2, \text{where}$$

$$\hat{f} = \frac{1}{n}\sum_{i=1}^{n} K_h(X - x_i).$$

x_i is observed data sampled from the underlying distribution, f, of a random variable X. The parameters $K(.)$ and h represent the kernel and the bandwidth, respectively, for the kernel density estimation $\hat{f}$. n is the number of samples. The optimal value of the bandwidth parameter, h_{opt}, that minimizes the $MISE(\hat{f})$ can be derived [6] to be as follows:

$$h_{opt} = k_2^{-2/5}\left\{\int K(t)^2 dt\right\}^{1/5}\left\{\int f''(x)^2 dx\right\}^{-1/5} n^{-1/5}. \quad (1)$$

$K(t)$ represents a symmetric kernel with unit area centered at 0 , e.g, uniform, triangle, Epanechnikov. The second moment of the kernel, $K(t)$, is denoted by k_2. The expression for h_{opt} cannot be directly evaluated since it contains the term with the second derivative of the unknown underlying distribution function, $f''(x)$. In the rule of thumb approach, the unknown density function is chosen from a family of parametric density functions [1]. Motivated by [2], [6], and a recent work [3], we choose the underlying parametric density to be a standard normal density function. When $f(x)$ is replaced with the standard normal distribution, the h_{opt} can be written as follows:

$$h_{opt} = k_2^{-2/5} \left\{ \int K(t)^2 dt \right\}^{1/5} \left\{ \frac{3}{8} \pi^{-1/2} \sigma^{-5} \right\}^{-1/5} n^{-1/5}. \tag{2}$$

In Eq. (2), h_{opt} can be estimated for a particular choice of kernel, $K(t)$, and approximating the σ with the spread of observed data. In Silverman's rule of thumb, $K(t)$ is chosen as a Gaussian kernel. Table 1 shows the evaluation of the quantities, k_2 and $\int K(t)^2 dt$ for various kernels. In our experiments, we refer to Table 1 for computing h_{opt} in Eq. (2). We estimate the data spread by finding the minimum of sample standard deviation and interquartile range scaled by the factor of (1/1.34) [6].

Kernel	$K(t)$	$\int t^2 K(t) dt$	$\int K(t)^2 dt$
Uniform	$\frac{1}{2}\mathbf{1}_{\{\lvert t\rvert \leq 1\}}$	$\frac{1}{3}$	$\frac{1}{2}$
Triangle	$(1-\lvert t\rvert)\mathbf{1}_{\{\lvert t\rvert \leq 1\}}$	$\frac{1}{6}$	$\frac{2}{3}$
Epanechnikov	$\frac{3}{4}(1-t^2)\mathbf{1}_{\{\lvert t\rvert \leq 1\}}$	$\frac{1}{5}$	$\frac{3}{5}$
Gaussian	$\frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}t^2}$	1	$\frac{1}{2\sqrt{\pi}}$

Table 1

The rule of thumb approach, however, holds a strong assumption about the underlying density to be from a parametric family. In the plugin approach [1], the unknown quantity $\int f''(x)^2 dx$ in Eq. (1) is estimated using nonparametric density estimation instead of choosing f from a family of

parametric density functions. This approach falls under the second generation methods and provides stable and consistent results when compared to first generation methods, e.g., rule of thumb approach [1]. We use the package 'KernSmooth' of R project for estimating bandwidth using the plugin method [5]. For the nonlocal statistics technique for density estimation described in Subsection 3.3 in the paper, each kernel of the nonparametric density estimation is weighed unequally. Wang et al. [7] show that the expression for the optimal bandwidth cannot be derived in closed form for weighted kernel density estimation using a plugin method [5]. Therefore, we estimate the bandwidth using the rule of thumb approach in nonlocal statistics technique for density estimation.

The most suitable choice of the kernel, $K(t)$, that minimizes the MISE can be derived to be an Epanechnikov kernel [6]. However, though the choice of kernel is important for the nonparametric density estimation, its influence on the estimated density is not as strong as the bandwidth selection when MISE is used as the error measure [6]. In our work, we present the results for uniform, triangle, and Epanechnikov kernels.

3 Model Complexity

The approaches described in Subsection 3.2 in the paper analytically compute the ratio distribution by going through each pair of kernels of pdf_X and pdf_Y. The computations take $O(nm)$, where n is number of kernels in pdf_X, and m is number of kernels in pdf_Y. Thus, the closed-form computation of ratio distribution has quadratic complexity for nonparametric density models. The ratio distribution can also be estimated empirically by Monte-Carlo sampling as discussed by Pöthkow and Hege [3]. However, when the underlying distributions are complex (e.g., with m and n kernels) the Monte-Carlo method [3] also needs heavy sampling to provide a reliable estimate of the probability distribution which is computed exactly by our method described in Subsection 3.2 in the paper. To assess the advantages of our closed-form solution, we compare the computational cost versus the accuracy offered by the sampling approach [3].

Fig. 2 shows the comparison of the efficiency of analytic and empirical approaches. The synthetic sphere dataset with a grid resolution of $64\times64\times64$ is injected with noise. At each grid point, 10 noisy samples are generated. The expected isosurface is extracted from this uncertain field. For extracting

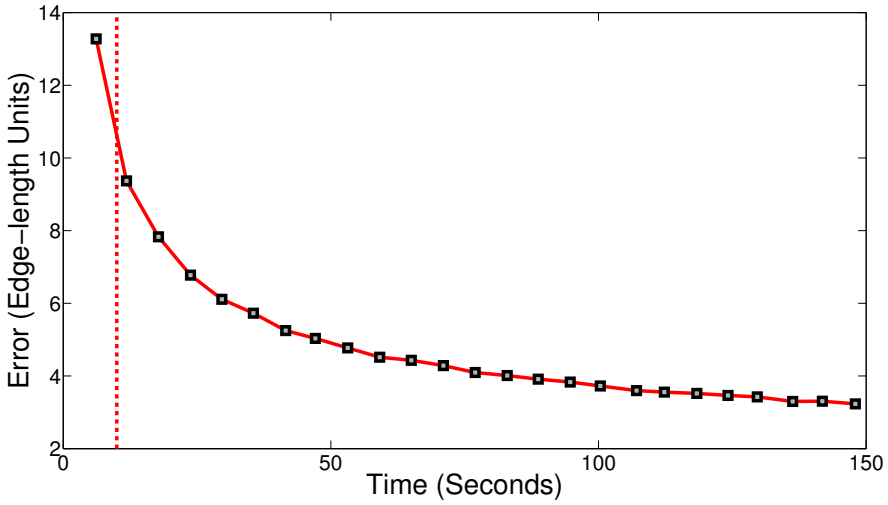

Figure 2: Timing comparison of analytic and sampling approaches [3] for the expected surface computation. The empirical error for the sampling approach (gray markers), indicated on the y-axis, versus the computation time on the x-axis. As sampling rates increase (by 5000 in each step), so does the computation time, and the error decreases. The red dotted vertical line represents the time required for the analytic solution (i.e., no-error). Clearly, the computational cost of the analytic solution is less than the sampling solution with 10000 samples whose error is around 10. Sampling with 5000 points is slightly faster, but with the error of about 15 edge-lengths.

the expected isosurface, the ratio density is computed, analytically and empirically, on each grid edge that is crossed by the isosurface. The expected isosurface extracted using the analytic approach takes 10 seconds on average which is indicated by the dotted vertical red line in Fig. 2. For the sampling approach [3], the timing results are shown for extracting the expected surface as the rate of sampling is increased. Starting with 5000 samples, the samples were increased by 5000 for 25 iterations. The error in the empirically extracted isosurface is computed by summing up the expectation error (in edge-length units) on each grid edge over the entire isosurface. Fig. 2 shows a plot of this error versus time. The error in expected surface computation drops down as the sampling is increased. Thus, in order for the empirical expected surface to converge to the analytic surface, it will require heavy Monte-Carlo sampling, and hence, high computational time.

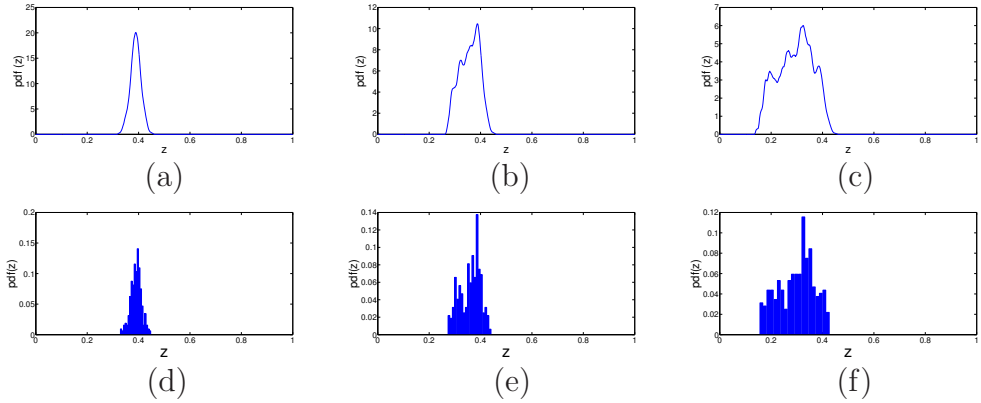

Figure 3: Effect of increasing complexity of underlying data distribution on the shape of empirical ratio density. (a) analytic ratio density for unimodal data distribution, (b) analytic ratio density for bimodal data distribution, (c) analytic ratio density for trimodal data distribution, (d) empirical ratio density for unimodal data distribution, (e) empirical ratio density for bimodal data distribution, (f) empirical ratio density for trimodal data distribution. The number of Monte-Carlo samples for empirical results is fixed to 320. The analytical results show that the shape of ratio density becomes more complex as the complexity of the underlying distribution grows from unimodal to trimodal. The empirical ratio density estimate is quite smooth and matches closely to analytic result when underlying density is unimodal. However, for bimodal and trimodal underlying distributions, the empirical ratio density estimate fluctuates and deviates more from corresponding analytical results. Thus, as the complexity of underlying density grows, higher Monte-Carlo sampling, and hence, higher computational time is required for reliable empirical estimates.

As the number of kernels used for the kernel density estimation is increased, the analytic time required to compute the ratio distribution increases quadratically. However, the higher number of kernels describe complex distributions, e.g, multimodal, skewed. To reliably estimate these complex distributions, the Monte-Carlo method requires extended sampling. Next, we show the impact of the complexity of underlying density on the sampling required for estimating the ratio density.

Fig. 3 shows the results of analytic and sampling-based estimation of the ratio density computed over an edge for various density functions with dif-

ferent complexities at edge vertices $\mathbf{v}_1$ and $\mathbf{v}_2$. The density functions at the edge vertices are designed with uniform kernels. Subfigures (a), (b), and (c) show the analytically computed ratio density functions for unimodal, bimodal, and trimodal distributions. Subfigures (d), (e), and (f) show the results of the Monte Carlo sampling approach corresponding to subfigures (a), (b), and (c), respectively. The rise in shape complexity of ratio density can be clearly seen in the analytic results as underlying density is changed from unimodal to trimodal density function. The number of samples in the Monte Carlo approach for all empirical results is fixed to 320. The shape of the empirical ratio density is reasonably smooth and matches closely with analytic density when underlying density is unimodal. However, with a rise in modality of underlying density, the empirical results start fluctuating and deviating from corresponding analytic density results. Thus, for reliable empirical ratio density estimates, higher Monte-Carlo sampling is required as the shape of the underlying density becomes more complex. An increase in the amount of Monte-Carlo sampling leads to higher computational cost. Therefore, although the analytic computation time increases quadratically with an increase in the number of kernels, the amount of Monte-Carlo sampling needs to be increased for reliable empirical results.

References

[1] M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91:401–407, 1996.

[2] Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[3] Kai Pöthkow and Hans Christian Hege. Nonparametric models for uncertainty visualization. 32(3.2):131–140, 2013.

[4] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[5] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53:683–690, 1991.

[6] B. W. Silverman. *Density estimation for statistics and data analysis.* Chapman & Hall/CRC, 1992.

[7] Bin Wang and Xiaofeng Wang. Bandwidth selection for weighted kernel density estimation. *Electronic Journal of Statistics*, 0:0–21, 2007.