Importance sampling performance metrics

1 IS coefficient of variation of the membership probability

The IS coefficient of variation of the membership probability, $\epsilon_P^{IS}(\tilde{n}, \Gamma_\ell, j)$, is an estimator of $\frac{\sqrt{\mathbb{V}\left(\hat{P}_{\tilde{n}}(\Gamma_\ell,j,y)\right)}}{\mathbb{E}\left(\hat{P}_{\tilde{n}}(\Gamma_\ell,j,y)\right)}$, with $\hat{P}_n(\Gamma_\ell, j, y) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{y(\tilde{X}^k)\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}_k)}$. It is computed for different quantizations Γ_ℓ^r in the family $(\Gamma_\ell^r)_{r\in\{1,\dots,n_\Gamma\}}$. These n_{Γ_ℓ} sets of prototypes Γ_ℓ^r are random perturbations of the prototypes from the true simulations: we first perform Lloyd's algorithm with the maps from the hydraulic simulators and the n_{Γ_ℓ} quantizations are created by applying to every pixel of the ℓ obtained maps a random change of between -20% and $+20\%$ of its magnitude.

We have $\mathbb{V}\left(\hat{P}_{\tilde{n}}(\Gamma_\ell, j, y)\right) = \frac{1}{\tilde{n}}\mathbb{V}\left(\mathbb{1}_{y(\tilde{X})\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X})}{g(\tilde{X})}\right)$. Two methods are now given to estimate this variance.

1.1 With a large input sample (analytical case)

Considering $(\tilde{X}^k)_{k=1}^{n_v}$ a large number of i.i.d random variable of density ν, an empirical estimator of $\mathbb{V}\left(\mathbb{1}_{y(\tilde{X})\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X})}{g(\tilde{X})}\right)$ is,

$$\frac{1}{n_v - 1}\sum_{k=1}^{n_v}\left[\mathbb{1}_{y(\tilde{X}^k)\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}^k)} - \hat{P}_{n_v}(\Gamma_\ell, j, y)\right]^2 .$$

Then an estimator of $\frac{\sqrt{\mathbb{V}\left(\hat{P}_{\tilde{n}}(\Gamma_\ell,j,y)\right)}}{\mathbb{E}\left(\hat{P}_{\tilde{n}}(\Gamma_\ell,j,y)\right)}$ is the IS relative standard deviation

$$\epsilon_P^{IS}(\tilde{n}, \Gamma_\ell^r, j) = \frac{1}{\sqrt{\tilde{n}}}\frac{\sqrt{\frac{1}{n_v-1}\sum_{k=1}^{n_v}\left[\mathbb{1}_{y(\tilde{X}^k)\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}^k)} - \hat{P}_{n_v}(\Gamma_\ell, j, y)\right]^2}}{\hat{P}_{n_v}(\Gamma_\ell, j, y)} .$$

1.2 Without a large input sample (coastal case)

If $y(\tilde{X}^k)$ cannot be computed for a large sample $(\tilde{X}^k)_{k=1}^{n_v}$, $\mathbb{V}\left(\mathbb{1}_{y(\tilde{X})\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X})}{g(\tilde{X})}\right)$ is estimated by a bootstrap approach, from the training database $(y(x^i))_{i=1,\dots,n_{\text{train}}}$, as the training inputs were sampled with density g as well.

The objective is to estimate $\mathbb{V}\left(\hat{P}_{1300}(\Gamma_\ell, j, y)\right)$ and then to deduce $\mathbb{V}\left(\mathbb{1}_{y(\tilde{X})\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X})}{g(\tilde{X})}\right) = 1300 \times \mathbb{V}\left(\hat{P}_{1300}(\Gamma_\ell, j, y)\right)$.

We proceed as follows:

1. First, we sample with replacement from the training inputs $(x^i)_{i=1,\dots,1300}$, leading to a set of new inputs denoted as $\tilde{X}^{\star(1)}_{(1)}, \dots, \tilde{X}^{\star(1300)}_{(1)}$
2. Compute $\hat{P}^{(1)}_{1300}(\Gamma_\ell, j, y) = \frac{1}{1300}\sum_{k=1}^{1300} \mathbb{1}_{y(\tilde{X}^{\star k}_{(1)})\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^{\star k}_{(1)})}{g(\tilde{X}^{\star k}_{(1)})}$ with this set of maps
3. Repeat this procedure 500 times by sampling with replacements to get $\hat{P}^{(i)}_{1300}(\Gamma_\ell, j, y) = \frac{1}{1300}\sum_{k=1}^{1300} \mathbb{1}_{y(\tilde{X}^{\star k}_{(i)})\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^{\star k}_{(i)})}{g(\tilde{X}^{\star k}_{(i)})}$ for $i \in 1, \dots, 500$ with different sets of 1300 maps sampled in the original training database.
4. Compute the empirical variance $\frac{1}{500-1}\sum_{i=1}^{500}(\hat{P}^{(i)}_{1300}(\Gamma_\ell, j, y) - \bar{P})^2$, with $\bar{P} = \frac{1}{500}\sum_{i=1}^{500} \hat{P}^{(i)}_{1300}(\Gamma_\ell, j, y)$
5. Multiply this empirical variance by 1300 to get an estimation of $\mathbb{V}\left(\mathbb{1}_{y(\tilde{X})\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X})}{g(\tilde{X})}\right)$, and deduce $\hat{\sigma}\left(\hat{P}_n(\Gamma_\ell, j, y)\right)$ the estimation of $\sigma\left(\hat{P}_n(\Gamma_\ell, j, y)\right) = \sqrt{\frac{1}{n}\mathbb{V}\left(\mathbb{1}_{y(\tilde{X})\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X})}{g(\tilde{X})}\right)}$
6. Compute $\epsilon_P^{IS}(\tilde{n}, \Gamma_\ell^r, j) = \frac{\hat{\sigma}\left(\hat{P}_n(\Gamma_\ell, j, y)\right)}{\hat{P}_{1300}(\Gamma_\ell, j, y)}$ the IS relative standard deviation.

2 IS centroid standard deviation

The IS centroid standard deviation, $\epsilon_{\Gamma_\ell}^{IS}(n_{\text{maps}}, \Gamma_\ell^r, j)$, is defined in relation with the variance $\mathbb{V}\left(\hat{E}_{n_{\text{maps}}}(\Gamma_\ell, j, y)\right)$. We have $(\hat{E}_{n_{\text{maps}}}(\Gamma_\ell, j, y))_i = \frac{\frac{1}{n_{\text{maps}}}\sum_{k=1}^{n_{\text{maps}}} y_{,i}(\tilde{X}^k)\mathbb{1}_{y(\tilde{X}^k)\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}^k)}}{\hat{P}_{n_{\text{maps}}}(\Gamma_\ell, j, y)}$ with $y_{,i}(\tilde{X}^k)$ the ith pixel of the map $y(\tilde{X}^k)$.

As Kempen and Van Vliet (2000) detail, the variance of the ratio of two random variables A and B can be approximated by the following expression:

$$\mathbb{V}\left(\frac{A}{B}\right) \approx \frac{\mu_A^2}{\mu_B^2}\left[\frac{\mathbb{V}(A)}{\mu_A^2} - 2\frac{\mathbb{C}\text{ov}(A, B)}{\mu_A \mu_B} + \frac{\mathbb{V}(B)}{\mu_B^2}\right] .$$

with μ_A, μ_B the mean of A and B.

In our case, we denote $A_i = \frac{1}{n_{\text{maps}}}\sum_{k=1}^{n_{\text{maps}}} y_{,i}(\tilde{X}^k)\mathbb{1}_{y(\tilde{X}^k)\in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}^k)}, i \in \{1, \dots, 64^2\}$, and $B = \hat{P}_{n_{\text{maps}}}(\Gamma_\ell, j, y)$

Then, with a large sample $(\tilde{X}^k)_{k=1}^{n_v}$ of $\tilde{X}$, we can compute the empirical estimation of the needed quantities:

- $\hat{\mu}_{A_i} = \frac{1}{n_v} \sum_{k=1}^{n_v} y_{,i}(\tilde{X}^k) \mathbb{1}_{y(\tilde{X}^k) \in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}^k)}$
- $\hat{\mathbb{V}}(A_i) = \frac{1}{n_{\mathrm{maps}}} \frac{1}{n_v - 1} \sum_{k=1}^{n_v} \left(y_{,i}(\tilde{X}^k) \mathbb{1}_{y(\tilde{X}^k) \in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}^k)} - \hat{\mu}_{A_i} \right)^2$
- $\hat{\mu}_B = \hat{P}_{n_v}(\Gamma_\ell, j, y) = \frac{1}{n_v} \sum_{k=1}^{n_v} \mathbb{1}_{y(\tilde{X}^k) \in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}^k)}$
- $\hat{\mathbb{V}}(B) = \frac{1}{n_{\mathrm{maps}}} \frac{1}{n_v - 1} \sum_{k=1}^{n_v} \left(\mathbb{1}_{y(\tilde{X}^k) \in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}^k)} - \hat{\mu}_B \right)^2$
- $c\hat{o}v(A_i, B) = \frac{1}{n_{\mathrm{maps}}} \frac{1}{n_v - 1} \sum_{k=1}^{n_v}$
  $\left(y_{,i}(\tilde{X}^k) \mathbb{1}_{y(\tilde{X}^k) \in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}^k)} - \hat{\mu}_{A_i} \right) \left(\mathbb{1}_{y(\tilde{X}^k) \in C_j^{\Gamma_\ell}} \frac{f_X(\tilde{X}^k)}{g(\tilde{X}^k)} - \hat{\mu}_B \right)$

The standard deviation can now be calculated at each pixel$i \in \{1 \dots 64^2\}$,

$$\hat{e}((\hat{E}_{n_{\mathrm{maps}}}(\Gamma_\ell, j, y))_i) = \sqrt{\hat{\mathbb{V}} \left((\hat{E}_{n_{\mathrm{maps}}}(\Gamma_\ell, j, y))_i \right)} \; .$$

Finally, for a given Voronoi cell $C_j^{\Gamma_\ell}$, the 90%-quantile over i is evaluated which yields the IS standard deviation $\epsilon_{\Gamma_\ell}^{IS}(n_{\mathrm{maps}}, \Gamma_\ell, j)$.

When n_v is too large to compute and store all $y(\tilde{X}^k)$, $k = 1, \dots, n_v$, the variance of the ratio is obtained from a bootstrap approach similar to the one described in Section 1.2.

References

Kempen, G. and L. Van Vliet (2000, 05). Mean and variance of ratio estimators used in fluorescence ratio imaging. *Cytometry 39*, 300–5.