AGATE: Adaptive Gray Area-based TEchnique to Cluster Virtual Machines with Similar Behavior

Claudia Canali, *Member, IEEE,* and Riccardo Lancellotti, *Member, IEEE*

APPENDIX A
ESTIMATION OF DISTANCE FROM CLUSTER CENTROIDS IN PRESENCE OF CLUSTERING ERRORS

As pointed out in Section 3.3, the probability distribution of VM distances from cluster centroids can be modeled as a bivariate Gaussian distribution with parameters μ_i, μ_j, σ_i, σ_j. However, the proposed technique assumes a perfect knowledge of cluster composition. This assumption is clearly an oversimplification because VMs belonging to cluster i that are closer to the centroid c_j of cluster j are mis-classified. As a consequence, the actual distribution of points for the estimation of the parameters follows a different distribution, represented in Figure 1.

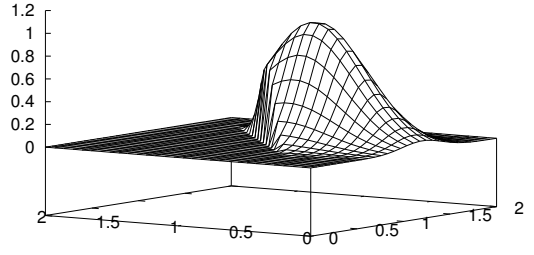

Fig. 1: Probability density of centroid distances

An attempt to use the output of the clustering phase for the estimation of the Gaussian parameters is most likely to lead to errors in the computation of $\epsilon_{j,i}$.

This section is devoted to the description of the statistical technique for a correct estimation the Gaussian parameters in the presence of mis-classified VMs.

For the sake of simplicity, we introduce for this analysis a simplified notation: let us consider two uncorrelated random variables X and Y. X is the random variable that describes the distance from centroid c_i, that is d_i, while Y describes the distance d_j. For consistency we rename the Gaussian distribution parameters as follows: $\mu_x = \mu_i$, $\mu_y = \mu_j$, $\sigma_x = \sigma_i$, $\sigma_y = \sigma_j$. This probability density of a bivariate Gaussian distribution can be described as the product of the two Gaussian functions for the un-correlated variables X and Y. However, we need to describe the probability density of the bivariate distribution subject to the bound that $y \leq x$. It is worth to note that the problem should include two additional constraints that are $x \geq 0$ and $y \geq 0$. However, in our case we can ignore this bound because the conditions $\mu_x > k\sigma_x$ and $\mu_x > k\sigma_x$, $k = 3$ typically apply. The resulting probability density is:

$$f(x,y) = \begin{cases} \frac{g(x,\mu_x,\sigma_x)\cdot g(y,\mu_y,\sigma_y)}{P(Y \leq X)}, & \text{if } y \leq x \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $g(\cdot,\mu,\sigma)$is the Gaussian function with mean μ and standard deviation σ, and $P(Y \leq X)$ is the probability that, given two samples x and y of the random variables X and Y, the condition $y \leq x$ is true.

We can compute $P(Y \leq X)$ as follows:

$$P(Y \leq X) = P(Y - X \leq 0)$$

but the difference of two Gaussian variables is another Gaussian variable with average equal to the sum of averages and standard deviation equal to the sum of standard deviations. Hence:

$$P(Y - X) \leq 0 = P[g(z, \mu_y - \mu_x, \sigma_x + \sigma_y) \leq 0] =$$
$$= P\left[g(z,0,1) \leq \frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right] = \Phi\left(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2 + \sigma_y^2}}\right)$$

where:

$$\Phi(t) = P[g(z,0,1) \leq t] = \int_{-\infty}^{t} \frac{e^{\frac{-\tau^2}{2}}}{\sqrt{2\pi}} d\tau$$

Having described the probability density of the distances from the centroids, we now need to estimate the parameters μ_x, μ_y, σ_x, σ_y given a set of samples $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$.

To this aim we exploit the principle of maximum likelihood, that is we aim to identify the parameters that maximize the likelihood function:

$$L\big(\mu_x,\mu_y,\sigma_x,\sigma_y;(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\big) =$$
$$= \prod_{i=1}^{n} f(x_i,y_i;\mu_x,\mu_y,\sigma_x,\sigma_y)$$

• *The authors are with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy.*
E-mail: claudia.canali@unimore.it,riccardo.lancellotti@unimore.it.

where $f(\cdot)$ is the probability density function in Equation 1. We consider more convenient to consider instead of the likelihood function $L(\cdot)$, its natural logarithm $\ln L(\cdot)$. That is, we aim to maximize:

$$\ln L(\cdot) = \ln\left(\frac{\prod_i g(x_i,\mu_x,\sigma_x)g(y_i,\mu_y,\sigma_y)}{\prod_i \Phi\left(\frac{\mu_x-\mu_y}{\sqrt{\sigma_x^2+\sigma_y^2}}\right)}\right) =$$

$$= \ln\left(\frac{\prod_i g(x_i,\mu_x,\sigma_x)\prod_i g(y_i,\mu_y,\sigma_y)}{\prod_i \Phi\left(\frac{\mu_x-\mu_y}{\sqrt{\sigma_x^2+\sigma_y^2}}\right)}\right) =$$

$$= \ln\left(\prod_i g(x_i,\mu_x,\sigma_x)\right) + \ln\left(\prod_i g(y_i,\mu_y,\sigma_y)\right) - n\ln\Phi\left(\frac{\mu_x-\mu_y}{\sqrt{\sigma_x^2+\sigma_y^2}}\right)$$

Finally, to identify the maximum of the likelihood function with respect to the four parameters μ_x, μ_y, σ_x, σ_y, we must find the points where the likelihood function derivative becomes zero. This determines the following system of four equation with four variables:

$$\begin{cases} \frac{\partial \ln L(\cdot)}{\partial \mu_x} = 0 \\ \frac{\partial \ln L(\cdot)}{\partial \mu_y} = 0 \\ \frac{\partial \ln L(\cdot)}{\partial \sigma_x} = 0 \\ \frac{\partial \ln L(\cdot)}{\partial \sigma_y} = 0 \end{cases}$$

that is:

$$\begin{cases} \frac{\sum_i (x_i-\mu_x)}{\sigma_x^2} - \frac{n}{\sqrt{2\pi}}\frac{e^{\frac{\Delta^2}{2}}}{\Phi(\Delta)} = 0 \\ \frac{\sum_i (y_i-\mu_y)}{\sigma_y^2} + \frac{n}{\sqrt{2\pi}}\frac{e^{\frac{\Delta^2}{2}}}{\Phi(\Delta)} = 0 \\ -\frac{n}{\sigma_x} + \frac{\sum_i (x_i-\mu_x)^2}{\sigma_x^3} + \frac{n}{2}\frac{e^{\frac{\Delta^2}{2}}}{\Phi(\Delta)}(\sigma_x^2+\sigma_y^2)^{\frac{3}{2}}(\mu_x-\mu_y)\sigma_x = 0 \\ -\frac{n}{\sigma_y} + \frac{\sum_i (y_i-\mu_y)^2}{\sigma_y^3} + \frac{n}{2}\frac{e^{\frac{\Delta^2}{2}}}{\Phi(\Delta)}(\sigma_x^2+\sigma_y^2)^{\frac{3}{2}}(\mu_x-\mu_y)\sigma_y = 0 \end{cases}$$

where $\Delta = \frac{\mu_x-\mu_y}{\sqrt{\sigma_x^2+\sigma_y^2}}$

The exponential and the $\Phi(\cdot)$ functions can be removed through a linear combination of the equations in the system as follows:

$$\begin{cases} \frac{\sum_i (x_i-\mu_x)}{\sigma_x^2} + \frac{\sum_i (y_i-\mu_y)}{\sigma_y^2} = 0 \\ \frac{n}{\sigma_x^2} - \frac{\sum_i (x_i-\mu_x)^2}{\sigma_x^4} - \frac{n}{\sigma_y^2} + \frac{\sum_i (y_i-\mu_y)^2}{\sigma_y^4} = 0 \\ \frac{\frac{\sum_i (x_i-\mu_x)^2}{\sigma_x^4} - \frac{n}{\sigma_x^2}}{(\sigma_x^2+\sigma_y^2)^{\frac{3}{2}}(\mu_x-\mu_y)} + \frac{\sqrt{2\pi}}{n}\frac{\sum_i (x_i-\mu_x)}{\sigma_x^2} = 0 \\ \frac{\frac{\sum_i (y_i-\mu_y)^2}{\sigma_y^4} - \frac{n}{\sigma_y^2}}{(\sigma_x^2+\sigma_y^2)^{\frac{3}{2}}(\mu_x-\mu_y)} - \frac{\sqrt{2\pi}}{n}\frac{\sum_i (y_i-\mu_y)}{\sigma_y^2} = 0 \end{cases}$$

thus leaving a non-linear system of equations that can can be solved using numerical approximations.

The values of $\mu_x = \mu_i$, $\mu_y = \mu_j$, $\sigma_x = \sigma_i$, $\sigma_y = \sigma_j$, can thus be used in the formulas of Section 3.3 to cope with the missing VMs that have been mis-classified by the clustering step in the estimation of $\epsilon_{j,i}$.

ACKNOWLEDGMENT

The authors would like to thank Prof. Marco Maioli for his contribution to the statistical model described in Appendix A.