SANDIA REPORT

SAND2015-9507
Unlimited Release
Printed October 2015

Formulas for Fast Computation of Divergence Statistics Applied to Quantitative Performance Analysis

Philippe Pébay, Janine Bennett

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2015-9507
Unlimited Release
Printed October 2015

Formulas for Fast Computation of Divergence Statistics Applied to Quantitative Performance Analysis

Philippe Pébay, Janine Bennett
Sandia National Laboratories
P.O. Box 969
Livermore, CA 94551, U.S.A.
pppebay,jcbenne@sandia.gov

Abstract

In an earlier report [PB15], we reported on the extension to the statistical analysis capability of the Visualization Tool Kit (VTK), which we developed for the calculation of divergence statistics, with the particular aim of providing quantitative means for HPC performance analysis, of which we provided an example as well as user's manual. However, we did not provide the mathematical foundations for this work.

In the current report, we fill this void with the complete derivation of the formulas which we used in the divergence statistics engine. This provides the foundations for future work which will aim at generalizing these formulas for more detailed HPC performance analysis.

Acknowledgments

The authors would like to thank who Jeremy Wilke (Sandia National Laboratories) who encouraged this work in the context of performance analysis for extreme-scale computing.

Contents

This page intentionally left blank

1 Introduction

In earlier work [PB15], we reported on a divergence statistics extension which we added to the Visualization Tool Kit (VTK), [Kit10], which we developed for the sake of quantifying, in a statistical manner akin to measuring a distance, between an observed empirical distribution and a theoretical, "ideal" one. The main motivation for this work was the performance assessment of High Performance Computing (HPC) performance analysis, performed either experimentally (i.e., using values of metrics measured on a real, live system performing an actual computation) or by simulation (for instance, using the Structural Simulation Toolkit SST [WK15] developed at Sandia National Laboratories).

In the aforementioned report, we focused on implementation details, specifically regarding how our implementation, in the form of the `vtkDivergenceStatistics` class, fits within the scalable, parallel statistics tool kit which we have previously developed [PTBM11].

However, although we illustrated the above with applications of the divergence statistics approach to SST simulation cases, we did not provide the mathematical foundations of our method, for [PB15] was mostly intended as a user's guide. It has come however to our attention that our method may lend itself to a broader class of problems than what we first anticipated. Therefore, the goal of the present report is to provide the full derivation of the formulas that are utilized by the `vtkDivergenceStatistics` engine.

2 Statistical Divergences

In this section, we first present a summary on the notion of *statistical divergence*. We then choose a set of 5 such divergences, some of which are semi-distances or even distances, exhibiting diverse properties, which we selected to experiment with our performance assessment method.

2.1 Definitions

The term *statistical divergence* describes a class of functions whose particular aim is to quantify the discrepancy between two distributions. Specifically, a divergence is a positive definite bivariate function with positive values, but it is not requested that they satisfy the symmetry axiom nor the triangle inequality: as such, they do cannot be called *distances*. This is the reason why the more general term *divergence* is used.

In particular, one interesting family of statistical divergences is that of *f-divergences*, which are defined between two distributions with respective probability densities p and q as:

$$(\cdot||\cdot) : (p,q) \longmapsto \int_{\mathbf{R}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x$$

where f is a real function, convex in $\mathbb{R}_+$ and such that $f(1) = 0$, cf. [Bas10] for more details. In the case of discrete probability distributions, which is that which is of interest to us here, where P and Q instead refer to probability mass functions, the definition becomes:

$$(\cdot||\cdot) : (P,Q) \longmapsto \sum_{x_i \in \mathscr{S}} Q(i) f\left(\frac{P(x_i)}{Q(x_i)}\right)$$

where $\mathscr{S}$ denotes the union of the supports (i.e., where probability is nonzero) of P and Q. Note that in the case where those two sample spaces do not exactly overlap, some terms in the sum become degenerate and the divergence must be calculated as a limit.

2.2 A Selection of Divergences with Diverse Properties

Infinitely many statistical divergences may be conceived with the above definition. However, there is a number of classical formulas, which are used in different contexts, that are well known. For example, the *Kullback-Leibler divergence* is widely used in the field of information theory, and is related to several concepts such as mutual information and Shannon entropy. Another classical example is that of the *Bhattacharyya semi-distance*, a measure of statistical overlap between two samples, which has many applications in computer vision, when attempting to match two different observations based on their respective color histograms.

Different statistical divergences and distances often provide qualitatively similar results but, as we have observed in our analyses, frequently reveal different details which might be of interest, or

not. Not knowing, *a priori*, which divergence would be best in general, or even whether one could be considered best in the context of HPC performance analysis, we retained 5 different divergences, selected amongst the "classic" ones commonly described in the literature. This choice is somewhat arbitrary ; however, it offers enough variety, from divergences in the narrowest sense of the term, to semi-distances or distances in the full meaning of term, so that this study appears to be the first in its kind to propose this approach with such relative generality.

Our selection hence goes as follows:

- The *total variation distance* $\delta(\cdot,\cdot)$, or TVD, obtained when $f(u) = \frac{1}{2}|u-1|$, and which is a distance in the true sense of the term (being symmetric and satisfying the triangle inequality). Note that in our case, where we focus on discrete distributions, it is the same distance as the *1-distance*, up to a factor of 2:

$$\delta(p,q) = \frac{1}{2}\|p-q\|_1 .$$

  We prefer to use the TVD for it has an upper bound of 1, which is convenient, but in statistical literature the 2 are sometimes confounded. In the case of discrete probability distributions, the above formula becomes:

$$\delta(P,Q) = \frac{1}{2}\sum_{x_i \in \mathscr{S}} |P(x_i) - Q(x_i)|$$

- The *Hellinger distance* $d_H(\cdot,\cdot)$, also symmetric and satisfying the triangle inequality, obtained by taking the square root of the f-divergence associated with $f(u) = \frac{1}{\sqrt{2}}(\sqrt{u}-1)^2$. Again in the case of discrete distribution, there is a relationship with a known norm, in this case the 2-distance (or Euclidean distance): specifically,

$$d_H(p,q) = \frac{1}{\sqrt{2}}\|\sqrt{p}-\sqrt{q}\|_2 .$$

  In addition, d_H also has an upper bound of 1. In the case of discrete distributions, one obtains.

$$d_H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i \in \mathscr{S}} \left(\sqrt{P(x_i)} - \sqrt{Q(x_i)}\right)^2}.$$

- The *Bhattacharyya coefficient* $b(\cdot,\cdot)$ is another statistical divergence which, albeit not a f-divergence (because the function f does not vanish at $u=1$), is obtained with $f(u) = \sqrt{u}$. We then define he *Bhattacharyya semi-distance* as follows :

$$d_B(p,q) = -\log b(p,q).$$

  In this case again, we are interested in the formula that arises for discrete distributions:

$$d_B(P,Q) = -\log \sum_{i \in \mathscr{S}} \sqrt{P(x_i)Q(x_i)}.$$

It is a semi-distance because it satisfies all the axioms of a distance, except for the triangle inequality. In particular, unlike divergences in general, it is symmetric. However it is not bounded above and takes on an infinite value when the respective supports of p and q are disjoint.

- The *Kullback-Leibler divergence* $\Delta(\cdot\|\cdot)$ is obtained with $f(u) = u\log_2 u$:

$$\Delta(p,q) = \int_{\mathbb{R}} p(x)\log_2\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x.$$

Note that the natural logarithm is also often encountered in the literature, instead of $\log_2$. In the discrete case, the formula becomes as follows:

$$\Delta(P\|Q) = \sum_{i\in\mathscr{S}} P(x_i)\log_2\frac{P(x_i)}{Q(x_i)}$$

Albeit not a statistical distance, because of its lacking symmetry, it is nonetheless very useful as is allows one to give different meanings to the two distributions, where the first one represents a "model", in the sense of desired outcome, against which the second distribution is compared. In an information-theoretical context, $\Delta(p\|q)$ quantifies the amount of information that is lost when q is observed instead of the "ideal" p. In the discrete and finite case often encountered, as is the case for our application domain, the binary logarithm allows for a direct, intuitive understanding of the divergence: specifically, if the model distribution consists of a single, ideal value, then a divergence equal to some integer value of $a \in \mathbb{N}^*$ indicates that only $\frac{1}{2^a}^{\text{th}}$ of the observed sample has the desired outcome. It is also worth noticing that the Kullback-Leibler divergence is not bounded above and that, when P and Q do not have exactly the same support, it can take its values in $[0,+\infty]$, as a result of the fact that

$$0\log 0 = \lim_{x\to 0^+} x\log x = 0$$

and

$$\frac{1}{\log 0} = \lim_{x\to 0^+}\frac{1}{\log x} = +\infty.$$

- The χ^2 *divergence* $\chi^2(\cdot\|\cdot)$, obtained with $f(u) = (u-1)^2$:

$$\Delta(p,q) = \int_{\mathbb{R}} \frac{(p(x)-q(x))^2}{q(x)}\mathrm{d}x.$$

Also called *Pearson divergence*, it plays an important role in statistical literature in particular as result of its relationship with the homonymous χ^2 hypothesis testing. It is also a divergence, in the most strict meaning of the term, neither being symmetric nor satisfying the triangle inequality. When the distributions are discrete, the formula becomes

$$\chi^2(Q\|P) = \sum_{i\in\mathscr{S}} \frac{(P(x_i)-Q(x_i))^2}{Q(x_i)}.$$

Using again the limits for degenerate cases when the denominator vanishes, $\chi^2(\cdot\|\cdot)$ takes its values in $[0,+\infty]$.

Note that, in the context of HPC performance analysis, performed either experimentally (i.e., measured values on a real, live system performing an actual computation) or by simulation (for instance, using Sandia's Structural Simulation Toolkit SST), typical analyses will have divergence analyses repeated at regular time intervals, in order to obtain a time-series analysis which can be further processed to obtain a space-time quantitative performance aggregate value. This time-series analysis is outside the scope of this report.

3 Divergence Statistics for Performance Assessment

We now describe how we use the statistical divergences chosen in §2 for the sake of HPC performance analysis.

Our approach is to compare an observed (i.e., empirical) distribution of values for some set of variables of interest (e.g., measurements of network traffic, CPU utilization, probe temperature, etc.) with respect to an ideal distribution, materialized by a probability mass function whose entire weight (1) is located at a user-defined, variable-specific "ideal" value. For example, in the case of CPU utilization, our method can be used to quantify the discrepancy between the empirical distribution observed across a number of compute cores with respect to an ideal 100% CPU load for all cores.

3.1 Statement of the Problem

Consider a variable of interest, whose experimental or simulated values across a finite domain of interest are regarded as the realizations of a discrete random variable X, with empirical probability mass function (EPMF) denoted P. The support of P is the set of values where P has non-zero value, i.e., correspond to the values really observed (at least once), in the experiment or the simulation. Denoting $N \in \mathbb{N}^*$ is the number of such distinct values, the support of P is:

$$\text{supp}(P) = \{x_i\}_{i=1}^{i=N}$$

and by definition of a probability mass function, one has $\sum_{i=1}^{i=N} P(x_i) = 1$.

Now, assume that a value, denoted x_0, is considered ideal, (or "peak"), for the same variable of interest. Note that x_0 does not necessarily belong to $\text{supp}(P)$. We henceforth denote Q the PMF of the discrete random variable which has a single outcome, x_0. In other words,

$$\text{supp}(Q) = \{x_0\}$$

and $Q(x_0) = 1$.

It is important to note that the realizations of the variable of interest may, or may not, contain the peak value. For some of the chosen divergences, the latter will result in infinite values, whereas the statistical distances are bounded.

3.2 Formulas

We now explicitly derive the formulas which we implemented in `vtkDivergenceStatistics`: rather than computing the formulas given in §3, across the entire support of P (which can be arbitrarily large), we have derived much simpler expressions in the case where the support of Q is a singleton, as is the case here.

Theorem 3.1. *With the above setting, and using the convention that* $-\log 0 = +\infty$ *together with the usual arithmetic operations on the extended real number line* $\mathbb{R} \cup \{-\infty; +\infty\}$*, then:*

$$\begin{aligned}
\delta(Q,P) &= 1 - P(x_0), \\
d_H(Q,P) &= \sqrt{1 - \sqrt{P(x_0)}}, \\
d_B(Q,P) &= -\log \sqrt{P(x_0)}, \\
\Delta(Q\|P) &= -\log_2 P(x_0), \\
\chi^2(Q\|P) &= \frac{1}{P(x_0)} - 1.
\end{aligned}$$

Proof. We prove this theorem by disjunction of cases. First, consider the case when $x_0 \notin \operatorname{supp}(P)$; therefore,

$$\begin{aligned}
\operatorname{supp}(P) \cup \operatorname{supp}(Q) &= \{x_i\}_{i=0}^{i=N}, \\
\forall i \in \mathbb{N}, 1 \leq i \leq N \;\; Q(x_i) &= 0,
\end{aligned}$$

and $P(x_0) = 0$. Using the arithmetic operations on the extended real number line $\mathbb{R} \cup \{-\infty; +\infty\}$, together with the limits of $\log x$ and $x \log x$ as $x \to 0^+$, we obtain the following identities:

$$\delta(Q,P) = \frac{1}{2}\|Q-P\|_1 = \frac{1}{2}\sum_{i=0}^{i=N} |Q(x_i) - P(x_i)| = \frac{1}{2}\Big(Q(x_0) + \sum_{i=1}^{i=N} P(x_i)\Big) = \frac{1}{2}(1+1) = 1,$$

$$d_H(Q,P) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=0}^{i=N} \left(\sqrt{Q(x_i)} - \sqrt{P(x_i)}\right)^2} = \frac{1}{\sqrt{2}}\sqrt{Q(x_0) + \sum_{i=1}^{i=N} P(x_i)} = \frac{\sqrt{2}}{\sqrt{2}} = 1,$$

$$d_B(Q,P) = -\log \sum_{i=0}^{i=N} \sqrt{Q(x_i)P(x_i)} = -\log\Big(\sqrt{Q(x_0) \times 0} + \sum_{i=1}^{i=N} \sqrt{0 \times P(x_i)}\Big) = +\infty,$$

$$\Delta(Q\|P) = \sum_{i=0}^{i=N} Q(x_i) \log_2 \frac{Q(x_i)}{P(x_i)} = 1 \times \log_2 \frac{1}{0} + \sum_{i=1}^{i=N} \frac{0 \log_2 0}{P(x_i)} = +\infty + 0 = +\infty,$$

$$\chi^2(Q\|P) = \sum_{i=0}^{i=N} \frac{(Q(x_i) - P(x_i))^2}{P(x_i)} = \frac{(1-0)^2}{0} + \sum_{i=1}^{i=N} P(x_i) = +\infty + 1 = +\infty.$$

On the other hand,

$$\begin{aligned}
1 - P(x_0) &= 1, \\
\sqrt{1 - \sqrt{P(x_0)}} &= 1, \\
-\log \sqrt{P(x_0)} &= +\infty, \\
-\log_2 P(x_0) &= +\infty, \\
\frac{1}{P(x_0)} - 1 &= +\infty,
\end{aligned}$$

which proves the identities when $x_0 \notin \operatorname{supp}(P)$.

Second, consider the case when $x_0 \in \mathrm{supp}(P)$; therefore,

$$\mathrm{supp}(P) \cup \mathrm{supp}(Q) = \mathrm{supp}(P) = \{x_i\}_{i=1}^{i=N},$$

which implies that $x_0 = x_q$, for some unique index q between 1 and N. Therefore, that $P(x_0) = P(x_q) \neq 0$ and $Q(x_0) = Q(x_q) = 1$. We can thus derive the desired identities, as follows:

$$\delta(Q,P) = \frac{1}{2}\sum_{i=1}^{i=N} |Q(x_i) - P(x_i)| = \frac{1}{2}\left(1 - P(x_q) + \sum_{\substack{i=1 \\ i\neq q}}^{i=N} P(x_i)\right) = \frac{1}{2}(2 - 2P(x_q)) = 1 - P(x_0),$$

$$d_H(Q,P) = \sqrt{\frac{\sum_{i=1}^{i=N}\left(\sqrt{Q(x_i)} - \sqrt{P(x_i)}\right)^2}{2}} = \sqrt{\frac{\left(1 - \sqrt{P(x_q)}\right)^2 + \sum_{\substack{i=1 \\ i\neq q}}^{i=N} P(x_i)}{2}} = \sqrt{1 - \sqrt{P(x_0)}},$$

$$d_B(Q,P) = -\log \sum_{i=1}^{i=N} \sqrt{Q(x_i)P(x_i)} = -\log\left(\sqrt{1 \times P(x_q)} + \sum_{\substack{i=1 \\ i\neq q}}^{i=N} \sqrt{0 \times P(x_i)}\right) = -\log\sqrt{P(x_0)},$$

$$\Delta(Q\|P) = \sum_{i=1}^{i=N} Q(x_i)\log_2\frac{Q(x_i)}{P(x_i)} = 1 \times \log_2\frac{1}{P(x_q)} + \sum_{\substack{i=1 \\ i\neq q}}^{i=N} \frac{0\log_2 0}{P(x_i)} = -\log_2 P(x_0),$$

$$\chi^2(Q\|P) = \sum_{i=1}^{i=N} \frac{(Q(x_i) - P(x_i))^2}{P(x_i)} = \frac{(1 - P(x_q))^2}{P(x_q)} + \sum_{\substack{i=1 \\ i\neq q}}^{i=N} P(x_i) = \frac{1}{P(x_q)} - 2 + \sum_{i=1}^{i=N} P(x_i) = \frac{1}{P(x_0)} - 1.$$

This completes the proof of the theorem, for no other case than either $x_0 \notin \mathrm{supp}(P)$ or $x_0 \in \mathrm{supp}(P)$ exists. □

Note that the respective roles of P and Q can be exchanged, leaving the results unchanged, for the total variation and Hellinger distances, as well as for the Bhattacharyya semi-distance, as a result of their satisfying the axiom of symmetry (being distances):

$$\begin{aligned}
\delta(P,Q) &= \delta(Q,P) = 1 - P(x_0), \\
d_H(P,Q) &= d_H(Q,P) = \sqrt{1 - \sqrt{P(x_0)}}, \\
d_B(P,Q) &= d_B(Q,P) = -\log\sqrt{P(x_0)}.
\end{aligned}$$

However, this is not true for any of the two strict divergences (Hellinger and χ^2), whose values are always equal to $+\infty$ if P and Q are permuted, even when $x_0 = x_q \in \mathrm{supp}(P)$, as shown below:

$$\Delta(P\|Q) = \sum_{i=1}^{i=N} P(x_i)\log_2\frac{P(x_i)}{Q(x_i)} = P(x_q)\log_2 P(x_q) + \sum_{\substack{i=1 \\ i\neq q}}^{i=N} \frac{P(x_i)\log_2 P(x_i)}{0} = +\infty,$$

$$\chi^2(P\|Q) = \sum_{i=1}^{i=N} \frac{(P(x_i) - Q(x_i))^2}{Q(x_i)} = (1 - P(x_q))^2 + \sum_{\substack{i=1 \\ i\neq q}}^{i=N} \frac{P(x_i)^2}{0} = +\infty.$$

References

[Bas10] M. Basseville. Divergence measures for statistical data processing. Publications Internes de l'IRISA PI-1961, IRISA, November 2010.

[Kit10] Inc. Kitware. *The VTK User's Guide, version 5.4*. Kitware, Inc., 2010.

[PB15] P. Pébay and J. Bennett. A divergence statistics extension to VTK for quantitative performance analysis. Sandia Report SAND2015-1152, Sandia National Laboratories, February 2015.

[PTBM11] P. Pébay, D. Thompson, J. Bennett, and A. Mascarenhas. Design and performance of a scalable, parallel statistics toolkit. In *Proc. 25th IEEE International Parallel & Distributed Processing Symposium, 12th International Workshop on Parallel and Distributed Scientific and Engineering Computing*, Anchorage, AK, U.S.A., May 2011.

[WK15] Jeremiah J. Wilke and Joseph P. Kenny. Using discrete event simulation for programming model exploration at extreme-scale: Macroscale components for the structural simulation toolkit (SST). Sandia Report SAND2015-1027, Sandia National Laboratories, February 2015.

DISTRIBUTION:

2	MS 9018	Central Technical Files, 8944
1	MS 0899	Technical Library, 9536


Sandia National Laboratories