Addendum: Statistical Evaluation of Fuzzing

In the following, we provide algorithm showing bootstrap-based methods for evaluating fuzzing evaluations.

Recall that our null hypothesis states that there is no discernible difference between a new technique and some baseline fuzzer. If the test statistics exceed a critical value, in fuzzing scenarios usually its 95% quantile, we reject the null hypothesis. Now, when using resampling tests, the critical value is given by the $1-\alpha$-quantile of a sample of simulated test statistics determined from so-called bootstrap samples, which are in our case sampled with replacement from the observed data and with the same sample size as for the latter.

We provide a *bootstrap version* of the two-sample t-test in Algorithms 1 and 2. If more than two fuzzers have been compared for a target, the two-sample t-test is not a good choice since we would have to perform more than one pairwise comparison to test the null hypotheses of no difference between any of the expected means for the fuzzing methods. This results in the *multiple testing problem*, which is the observation that the probability of at least one false positive result in the set of comparisons performed for a target exceeds the single test level α substantially. The same argument holds for other strategies based on two-sample comparisons, such as the Mann-Whitney-U test [1].

A solution to this problem is the bootstrap version of the *ANOVA method*. If the ANOVA rejects the no-effect hypothesis, it shows at level α that there is at least one pair of fuzzing methods that perform significantly differently for the target considered. The second step is then to perform a so-called *Posthoc*-test, to determine which pairwise comparisons are significant *given that the ANOVA has already shown that there are significant differences at all*. Possible *Posthoc*-tests are, for example, the Tukey-Kramer method if all pairwise comparisons among all samples are of interest or the Dunnett method if only the comparisons to a reference method, such as the newly developed fuzzer, are of interest [4]. To use bootstrap versions of these algorithms, a simple solution are two-sample t-tests with critical values for rejection based on bootstrap resampling of the test statistics. Here, one uses for each simulation the maximal value of test statistics for all pairwise comparisons of interest. Algorithms 3 and 4 summarize bootstrap versions of the ANOVA algorithm and Algorithm 5 shows a Posthoc-test for the comparison.

Statistical tests usually require a critical value c, where the test rejects the null hypothesis H_0, if the test statistics $T > c$. In a lot of statistical frameworks, the threshold c may be unknown, particularly, if the distribution of the data is

substantially different from a simple parametric distribution, and/or the sample size is small. A frequent solution is to estimate the true theoretical threshold c by a suitable bootstrap algorithm, which is based on simulating the test statistic from artificially generated random data with a distribution closely similar to the true, but unknown, distribution of the data. In more detail, the main concept of a bootstrap method is to simulate the distributional behavior of the data and test statistics under H_0 by repeated application of the test to artificial bootstrap datasets generated by some resampling procedure from the original data. Here, two numbers are of interest. First, the typical number of bootstrap replications, i.e. simulation runs, which can be chosen by the user, can e.g. be 1000. Second, the sample size of the original data sample is of interest since the bootstrap method is based on a sufficient approximation of the true distribution of the data by the sample. Moreover, the sample size determines the maximal number of different bootstrap samples which can be generated. A reasonable minimal sample size may e.g. be $8 - 10$ [3], which yields $\binom{15}{8} \dots \binom{19}{10} \approx 6400 \dots 92000$ different samples. Bootstrap methods have been shown to be effective in empirical studies (see e.g. [2]) and it was proven that they result in consistent decision rules (cf. [5]). The method we use is called empirical bootstrap (cf. [5]).

Here, we apply a resampling algorithm that depends on the data set and repeatedly resample this data with replacement and compute the value of the test statistics T from this data to imitate the behavior of T given H_0 to get an empirical estimator c^*. To this end, if the test rejects for $T > c$ and T is resampled exactly B times, we order the resulting values, $S_1^*, \dots, S_B^*$ and estimate c by $S(\lfloor (1-\alpha)B \rfloor)$.

Algorithm 1 Threshold Bootstrap for the t-test

Require: Measurements: $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$,
bootstrap iterations: B, type-1 error: α

Ensure: c.

function threshold_c(x, y, B, α)
  Join samples x and y:
  $z = (x_1, \ldots, , x_{n_x}, y_1, \ldots, y_{n_y})$
  Shift samples to satisfy H_0
  $x' = x - \bar{x} + \bar{z},\ y' = y - \bar{y} + \bar{z}$
  for $i = 1, \ldots, B$ **do**
    Sample $x^* = sample(x', n_x), y^* = sample(y', n_y)$
    Compute pooled standard deviation
    $S^2_{x^*} = \frac{1}{n_x} \sum_{i=1}^{n} \left(x_i^* - \bar{x}^*\right)^2$, $S^2_{y^*} = \frac{1}{n_y} \sum_{i=1}^{n} \left(y_i^* - \bar{y}^*\right)^2$.
    Compute test statistics for bootstrap data
    $T_i^*(x^*, y^*) = \frac{|\bar{x}^* - \bar{y}^*|}{\sqrt{\frac{S^2_{x^*}}{n_x} + \frac{S^2_{y^*}}{n_y}}}$
  end for
  Sort statistics in ascending order:
  $(T^*_{(1)}(x^*, y^*), ..., T^*_{(B)}(x^*, y^*)) = sort(T^*_{(1)}(x^*, y^*), ..., T^*_{(B)}(x^*, y^*))$.
  $c^* := T^*_{(\lfloor (1-\alpha)B \rfloor), j}(x^*)$.
  return c^*.
end function

Algorithm 2 Bootstrap t-test

Require: Measurements: $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$,
bootstrap iterations: B, type-1 error: α

Ensure: decision 0 or 1.

function test(x, y,B, α)
  c :=threshold_c(x, y, B, α)
  Compute pooled standard deviation
  $S_x^2 = \frac{1}{n_x} \sum_{i=1}^{n} (x_i - \bar{x})^2$, $S_y^2 = \frac{1}{n_y} \sum_{i=1}^{n} (y_i - \bar{y})^2$.
  Compute test statistics for data
  $T(x, y) = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}}$
  if $T \geq c$ **then**
    return 1
  else
    return 0
  end if
end function

Algorithm 3 Threshold Bootstrap for the ANOVA method

Require: Measurements: $x = (x_{i,1}, \ldots, x_{i,n_i}$ for samples $i = 1, \ldots, m$,
bootstrap iterations: B, type-1 error: α

Ensure: c.

function threshold_c-ANOVA($x_{i,j}$, $i = 1, \ldots, m$, $j = 1, \ldots, m_i$ B, α)
  Join samples
  $z = (x_{1,1}, \ldots, , x_{1,n_1}, \ldots, y_{m,1}, \ldots, y_{m,n_m})$
  Shift samples to satisfy H_0
  $x'_{i,} = x'_{i,} - \bar{x}_{i,} + \bar{z}$ $i = 1, \ldots, m$
  for $i = 1, \ldots, B$ **do**
    Sample $x^*_{1,} = sample(x'_{1,}, n_1), \ldots, x^*_{m,} = sample(x'_{m,}, n_m)$
    Compute test statistics for bootstrap data
    $T^*_i(x^*_{1,}, \ldots, x^*_{m,}) = F$
    where F is the test statistics of the one-way ANOVA
    For the posthoc test
    $T^{p,*}_i(x^*_{1,}, \ldots, x^*_{m,}) = max($
      $t.test\$statistics(x^*_{1,}, x^*_{2,}), \ldots,$
      $t.test\$statistics(x^*_{1,}, x^*_{m,}), \ldots,$
      $t.test\$statistics(x^*_{m-1,}, x^*_{m,})$
    $)$
    where t.test$statistics(,) is the test-statistics of the two-sample t-test.
  end for
  Sort Anova statistics in ascending order:
  $(T_{(1)}p, *(x^*_{1,}, \ldots, x^*_{m,}), ..., T^{p,*}_{(B)}(x^*_{1,}, \ldots, x^*_{m,}) = sort(T^*_{(1)}(x^*_{1,}, \ldots, x^*_{m,}), ..., T^*_{(B)}(x^*_{1,}, \ldots, x^*_{m,}))$.
  $c^* := T^*_{(\lfloor(1-\alpha)B\rfloor),j}(x^*_{1,}, \ldots, x^*_{m,})$.
  Sort posthoc-statistics in ascending order:
  $(T^*_{(1)}(x^*_{1,}, \ldots, x^*_{m,}), ..., T^*_{(B)}(x^*_{1,}, \ldots, x^*_{m,}) = sort(T^{p,*}_{(1)}(x^*_{1,}, \ldots, x^*_{m,}), ..., T^{p,*}_{(B)}(x^*_{1,}, \ldots, x^*_{m,}))$.
  $c^{p,*} := T^{p,*}_{(\lfloor(1-\alpha)B\rfloor),j}(x^{p,*}_{1,}, \ldots, x^*_{m,})$.
  return $(c^*, c^{p,*})$.
end function

Algorithm 4 Bootstrap test for the ANOVA method

Require: Measurements: $x = (x_{i,1}, \ldots, x_{i,n_i}$ for samples $i = 1, \ldots, m$,
bootstrap iterations: B, type-1 error: α

Ensure: decision 0 or 1.

function test($x_{1,}$, $\ldots x_{m,}$,B, α)
  c :=threshold_c-ANOVA($x_{i,j}$, $i = 1, \ldots, m$, $j = 1, \ldots, m_i$ B, α)[1]
  Compute test statistics for data
  $T_i(x_{1,}, \ldots, x_{m,}) = F$
  where F is the test statistics of the one-way ANOVA
  if $T \geq c$ **then**
    return 1
  else
    return 0
  end if
end function

Algorithm 5 Posthoc test for the ANOVA method

Require: Measurements: $x = (x_{i,1}, \ldots, x_{i,n_i}$ for samples $i = 1, \ldots, m$,
bootstrap iterations: B, type-1 error: α

Ensure: decision 0 or 1.

function test($x_{1,}$, $\ldots x_{m,}$,B, α)
  c^p :=threshold_c-ANOVA($x_{i,j}$, $i = 1, \ldots, m$, $j = 1, \ldots, m_i$ B, α)[2]
  Compute test statistics for data

  for $i, j \in 1, \ldots, m,\ i < j$ **do**
    $T^p_{i,j}(x_{1,}, \ldots, x_{m,}) = t.test\$statistic(x_{i,}, x_{j,})$
    where t.test$statistic is the test statistics of the two-sample t-test
    if $T^p_{i,j}(x_{1,}, \ldots, x_{m,}) \geq c^p$ **then**
      "Comparison of sample i vs. j significant"
    else
      "Comparison of sample i vs. j not significant"
    end if
  end for
end function

References

[1] A. Arcuri and L. Briand, "A Practical Guide for Using Statistical Tests to Assess Randomized Algorithms in Software Engineering," 2011.

[2] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[3] P. Hall, *The Bootstrap and Edgeworth Expansion (Springer Series in Statistics)*, 1st ed., ser. Springer Series in Statistics. New York, NY: Springer New York, 1992.

[4] L. Sachs, *Applied Statistics: A Handbook of Techniques*, 2nd ed., ser. Springer Series in Statistics. New York, NY: Springer New York, 1984.

[5] A. W. v. d. Vaart, *Asymptotic Statistics*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.