Simpler, Faster, and More Robust T-Test Based Leakage Detection

A. Adam Ding[1], Cong Chen[2](✉), and Thomas Eisenbarth[2]

[1] Northeastern University, Boston, MA, USA
a.ding@neu.edu
[2] Worcester Polytechnic Institute, Worcester, MA, USA
{cchen3,teisenbarth}@wpi.edu

Abstract. The TVLA procedure using the t-test has become a popular leakage detection method. To protect against environmental fluctuation in laboratory measurements, we propose a paired t-test to improve the standard procedure. We take advantage of statistical matched-pairs design to remove the environmental noise effect in leakage detection. Higher order leakage detection is further improved with a moving average method. We compare the proposed test with standard t-test on synthetic data and physical measurements. Our results show that the proposed tests are robust to environmental noise.

1 Motivation

More than 15 years after the proposal of DPA, standardized side channel leakage detection is still a topic of controversial discussion. While Common Criteria (CC) testing is an established process for highly security critical applications such as banking smart cards and passport ICs, the process is slow and costly. While appropriate for high-security applications, CC is too expensive and too slow to keep up with the innovation cycle of a myriad of new networked embedded products that are currently being deployed as the Internet of Things. As a result, an increasing part of the world we live in will be monitored and controlled by embedded computing platforms that, without the right requirements in place, will be vulnerable to even the most basic physical attacks such as straightforward DPA.

A one-size-fits-most leakage detection test that is usable by non-experts and can reliably distinguish reasonably-well protected cryptographic implementations from insecure ones could remedy this problem. Such a test would allow industry to self-test their solutions and hopefully result in a much broader deployment of appropriately protected embedded consumer devices. The TVLA test was proposed as such a leakage detection test in [6,10]. The TVLA test checks if an application behaves differently under two differing inputs, e.g. one fixed input vs. one random input. As the original DPA, it uses averaging over a large set of observations to detect even most nimble differences in behavior, which can potentially be exploited by an attacker.

© Springer International Publishing Switzerland 2016
F.-X. Standaert and E. Oswald (Eds.): COSADE 2016, LNCS 9689, pp. 163–183, 2016.
DOI: 10.1007/978-3-319-43283-0_10

Due to its simplicity, it is applicable to a fairly wide range of cryptographic implementations. In fact, academics have started to adopt this test to provide evidence of existing leakages or their absence [1,3–5,13,15,16,20]. With increased popularity, scrutiny of the TVLA test has also increased. Mather et al. [14] studied the statistical power and computation complexity of the t-test versus mutual information (MI) test, and found that t-test does better in the majority of cases. Schneider and Moradi [19] for example showed how the t-test higher order moments can be computed in a single pass. They also discussed the tests sensitivity to the measurement setup and proposed a randomized measurement order. Durveaux and Standaert [8] evaluate the convenience of the TVLA test for detecting relevant points in a leakage trace. They also uncover the implications of good and bad choices of the fixed case for the fixed-vs-random version of the TVLA test and discuss the potential of a fixed-vs-fixed scenario.

However, there are other issues besides the choice of the fixed input and the measurement setup that can negatively impact the outcome for the t-test based leakage detection. Environmental effects can influence the t-test in a negative way, i.e., will decrease its sensitivity. In the worst case, this means that a leaky device may pass the test only because the environmental noise was strong enough. This is a problem for the proposed objective of the TVLA test, i.e. self-certification by non-professionals who are not required to have a broad background in side channel analysis.

Our Contribution. In this work, we propose the adoption of the paired t-test for leakage detection, especially in cases where long measurement campaigns are performed to identify nimble leakages. We discuss several practical issues of the classic t-test used in leakage detection and show that many of them can be avoided when using the paired t-test. To reap the benefits of the locality of the individual differences of the paired t-test in the higher order case, we further propose to replace the centered moments with a local approximation. These approximated central moments are computed over a small and local moving window, making the entire process a single-pass analysis. In summary, we show that

- the paired t-test is more robust to environmental noise such as temperature changes and drifts often observed in longer measurement campaigns, resulting in a faster and more reliable leakage detection.
- using moving averages instead of a central average results in much better performance for higher order and multivariate leakage detection if common measurement noise between the two classes of traces is present, while introducing a vanishingly small inaccuracy if no such common noise appears. The improvement of the moving averages applies both to the paired and unpaired t-tests.

In summary, we advocate the adoption of the paired t-test based on moving averages as a replacement of Welch' t-test for detecting leakages, as results are at least on par with the prevailing methodology while showing much better results in the presence of a common noise source.

2 Background

In the framework of [10], the potential leakage for a device under test (DUT) can be detected by comparing two sets of measurements $\mathcal{L}_A$ and $\mathcal{L}_B$ on the DUT. A popular test for the comparison is Welch's t-test, which aims to detect the mean differences between the two sets of measurements. The null hypothesis is that the two samples come from the same population so that their population means μ_A and μ_B are the same. Let $\bar{L}_A$ and $\bar{L}_B$ denote their sample means, s_A^2 and s_B^2 denote their sample variance, n_A and n_B denote the number of measurements in each set. Then the t-test statistic and its degree of freedom are given by

$$t_u = \frac{\bar{L}_A - \bar{L}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}, \qquad v = \frac{(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B})^2}{\frac{(\frac{s_A^2}{n_A})^2}{n_A - 1} + \frac{(\frac{s_B^2}{n_B})^2}{n_B - 1}}. \tag{1}$$

The p-value of the t-test is calculated as the probability, under a t-distribution with v degree of freedom, that the random variable exceeds observed statistic $|t_u|$. This is readily done in Matlab as $2 * (1 - \mathit{tcdf}(\cdot, v))$ and in R as $2 * (1 - \mathit{qt}(\cdot, \mathit{df} = v))$. The null hypothesis of no leakage is rejected when the p-value is smaller than a threshold, or equivalently when the t-test statistic $|t_u|$ exceeds a corresponding threshold. The rejection criterion of $|t_u| > 4.5$ is often used [10,19]. Since $Pr(|t_{df=v>1000}| > 4.5) < 0.00001$, this threshold leads to a confidence level > 0.99999.

For leakage detection, a *specific* t-test use two sets $\mathcal{L}_A$ and $\mathcal{L}_B$ corresponding to different values of an intermediate variable: $V = v_A$ and $V = v_B$. To avoid the dependence on the intermediate value and the power model, *non-specific* t-test often uses the *fixed versus random* setup. That is, the first set $\mathcal{L}_A$ is collected with a fixed plaintext x_A, while the second set $\mathcal{L}_B$ is collected with random plaintexts x_B drawn from the uniform distribution. Then if there is leakage through an (unspecified) intermediate variable V, then

$$L_A = V(k, x_A) + r_A \qquad L_B = V(k, x_B) + r_B, \tag{2}$$

where k is the secret key, r_A and r_B are random measurement noises with zero means and variance σ_A^2 and σ_B^2 respectively. The non-specific t-test can detect the leakage, with large numbers of measurements n_A and n_B, when the fixed intermediate state $V(k, x_A)$ differs from the expected value of the random intermediate state $E_{x_B}[V(k, x_B)]$ where the expectation is taken over the uniform random plaintexts x_B.

The power model is very general for t-test framework of [10]. The intermediate variable can be of various sizes, including one bit or one byte intermediate state. Particularly, the tester does not need to know the underlying power model for the unspecified t-test. The power model in most of the paper is kept abstract and general. The theory does not depend on any specific power model. We only specify the exact power model in simulation studies that generated the data.

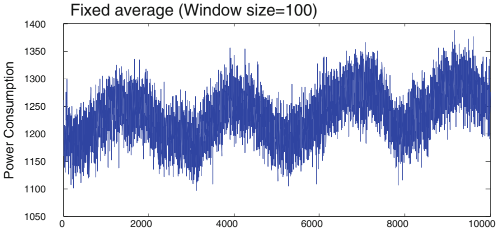


Fig. 1. Power consumption moving averages at a key-sensitive leakage point on the DPAv2 template traces

3 Methodology

This section introduces paired t-test and shows its superiority in a leakage model with environmental noise. The paired t-test retains its advantage of being a straightforward one-pass algorithm by making use of *moving* or local averages. By relying on the difference of matched pairs, the method is inherently numerically stable while retaining computational efficiency and parallelizability of the original t-test.

3.1 Paired T-Test

Welch's t-test works well when the measurement noises r_A and r_B are independent between the two sets of measurements. However, two sets of measurements can also share common variation sources during a measurement campaign. For example, power consumption and variance may change due to common environmental factors such as room temperature. While these environmental factors usually change slowly, such noise variation is more pronounced over a longer time period. With hard to detect leakages, often hundreds of thousands to millions of measurement traces are required for detection. These measurements usually take many hours and the environmental fluctuation is of concern in such situations. For example, for the DPA V2 contest, there are one million template traces collected over 3 days and 19 h, which show a clear temporal pattern [11]. Figure 1 (a subgraph of 2 in [11]) shows the average power consumption at 2373-th time point on the traces of DPAv2, using mean values over 100 non-overlapping subsequent traces.

Testing labs usually try to control the environmental factors to reduce such temporal variation. However, such effort can be expensive and there is no guarantee that all noise induced by environmental factors can be removed. Instead, we can deal with these environmental noise through statistical design. Particularly, we can adopt the *matched-pairs design* (Sect. 15.3 in [12]), where the measurements are taken in pairs with one each from the groups A and B. Then a *paired t-test* can be applied on such measurements, replacing the unpaired t-test (1). With n such pairs of measurements, we have n difference measurements $D = L_A - L_B$. The paired difference cancels the noise variation from the common source, making it easier to detect nonzero population difference. The null hypothesis of $\mu_A = \mu_B$ is equivalent to that the mean difference $\mu_D = 0$, which is tested by a paired t-test. Let $\bar{D}$ and s_D^2 denote the sample mean and sample variances of the paired differences D_1, ..., D_n. The paired t-test statistic is

$$t_p = \frac{\bar{D}}{\sqrt{\frac{s_D^2}{n}}}, \tag{3}$$

with the degree of freedom $n-1$. The null hypothesis of non-leakage is rejected when $|t_p|$ exceeds the threshold of 4.5.

To quantify the difference between the two versions of t-test, we can compare the paired t-test (3) and the unpaired t-test (1) here with $n_A = n_B = n$.

First, without common variation sources under model (2), $Var(D) = Var(L_A) + Var(L_B) = \tilde{\sigma}_A^2 + \tilde{\sigma}_B^2$. Here $\tilde{\sigma}_A^2 = \sigma_A^2 + Var[V(k, x_A)]$ and $\tilde{\sigma}_B^2 = \sigma_B^2 + Var[V(k, x_B)]$. Notice that $\bar{D} = \bar{L}_A - \bar{L}_B$, so for large n, the paired t-test and unpaired t-test are equivalent with $t_u \approx t_p \approx (\bar{L}_A - \bar{L}_B)/\sqrt{(\tilde{\sigma}_A^2 + \tilde{\sigma}_B^2)/n}$. The paired t-test works even if the two group variances are unequal $\tilde{\sigma}_A^2 \neq \tilde{\sigma}_B^2$. The two versions of the t-test perform almost the same in this case.

However, the paired t-test detects leakage faster if there are common noise variation sources. To see this, we explicitly model the common environmental factor induced variation not covered by model (2).

$$L_A = V(k, x_A) + r_A + r_E \qquad\qquad L_B = V(k, x_B) + r_B + r_E, \tag{4}$$

where r_E is the noise caused by common environmental factors, with mean zero and variance σ_E. The r_A and r_B here denote the random measurement noises excluding common variations so that r_A and r_B are independent, with zero means and variance σ_A^2 and σ_B^2 respectively. Again we denote $\tilde{\sigma}_A^2 = \sigma_A^2 + Var[V(k, x_A)]$ and $\tilde{\sigma}_B^2 = \sigma_B^2 + Var[V(k, x_B)]$. Then $t_u \approx (\bar{L}_A - \bar{L}_B)/\sqrt{(\tilde{\sigma}_A^2 + \tilde{\sigma}_B^2 + 2\sigma_E^2)/n}$ while $t_p \approx (\bar{L}_A - \bar{L}_B)/\sqrt{(\tilde{\sigma}_A^2 + \tilde{\sigma}_B^2)/n}$. The paired t-test statistic $|t_p|$ has a bigger value than the unpaired t-test $|t_u|$, thus identifies the leakage more efficiently. The difference increases when the environmental noise σ_E increases. Hence, the paired t-test performs as well or better than the unpaired test. However, the matched-pairs design of the paired t-test cancels common noise found in both pairs, making the test more robust to suboptimal measurement setups and environmental noise.

3.2 Higher Order and Multivariate Leakage Detection

The t-test can also be applied to detect higher order leakage and multivariate leakage [10,19]. For d-th order leakage at a single time point, the t-test compares sample means of $(L_A - \bar{L}_A)^d$ and $(L_B - \bar{L}_B)^d$. Under the matched-pairs design, the paired t-test would simply work on the difference

$$D = [(L_A - \bar{L}_A)^d - (L_B - \bar{L}_B)^d] \tag{5}$$

to yield the test statistic (3): $t_p = \bar{D}/\sqrt{s_D^2/n}$. Multivariate leakage combines leakage observation at multiple time points. A d-variate leakage combines leakage $L^{(1)}$, ..., $L^{(d)}$ at the d time points t_1, ..., t_d respectively. The combination is done through the centered product $CP(L^{(1)}, ..., L^{(d)}) = (L^{(1)} - \bar{L}^{(1)})(L^{(2)} - \bar{L}^{(2)}) \cdots (L^{(d)} - \bar{L}^{(d)})$. The standard d-variate leakage detection t-test compares the sample means of $CP(L_A^{(1)}, ..., L_A^{(d)})$ and $CP(L_B^{(1)}, ..., L_B^{(d)})$ with statistic (1). The paired t-test (3) uses the difference $D = [CP(L_A^{(1)}, ..., L_A^{(d)}) - CP(L_B^{(1)}, ..., L_B^{(d)})]$.

However, these tests (including the paired t-test) do not eliminate environmental noise effects on the higher order and multivariate leakage detection. The centering terms (the subtracted $\bar{L}$) in the combination function also need adjustment due to environmental noises, which are not random noise but follow some temporal patterns. To see this, we use the bivariate leakage model for first-order masked device as an example.

The leakage measurements at the two time points t_1 and t_2 leak two intermediate values $V^{(1)}(k, x, m)$ and $V^{(2)}(k, x, m)$ where k, x and m are the secret key, plaintext and mask respectively. For uniformly distributed m, $V^{(1)}(k, x, m)$ and $V^{(2)}(k, x, m)$ both follow a distribution not affected by k and x, therefore no first order leakage exits. Without loss of generality, we assume that $E_m[V^{(1)}(k, x, m)] = E_m[V^{(2)}(k, x, m)] = 0$, and the second order leakage comes from the product combination $V^{(1)}V^{(2)}$. [18] derived the strongest leakage combination function under a second order leakage model without the environmental noises:

$$L^{(1)} = c^{(1)} + V^{(1)}(k, x, m) + r^{(1)}, \qquad L^{(2)} = c^{(2)} + V^{(2)}(k, x, m) + r^{(2)}, \quad (6)$$

where $r^{(1)}$ and $r^{(2)}$ are zero-mean random pure measurement noises with variance σ_1^2 and σ_2^2 respectively. Under model (6), [18] showed that centered product leakage $(L^{(1)} - c^{(1)})(L^{(2)} - c^{(2)})$ is the strongest. Since c_1 and c_2 are unknown in practice, they are estimated by $\bar{L}^{(1)} = \bar{c}^{(1)} + \bar{V}^{(1)} + \bar{r}^{(1)}$ and $\bar{L}^{(2)} = \bar{c}^{(2)} + \bar{V}^{(2)} + \bar{r}^{(2)}$. With large number of traces, $\bar{L}^{(1)} \approx \bar{c}^{(1)}$ and $\bar{L}^{(2)} \approx \bar{c}^{(2)}$ by the law of large number. Hence $(L^{(1)} - \bar{L}^{(1)})(L^{(2)} - \bar{L}^{(2)})$ approximate the optimal leakage $(L^{(1)} - c^{(1)})(L^{(2)} - c^{(2)})$ well. However, considering environment induced noises, this is no longer the strongest leakage combination function. Let us assume that

$$L^{(1)} = c^{(1)} + V^{(1)}(k, x, m) + r^{(1)} + r_E^{(1)}, \; L^{(2)} = c^{(2)} + V^{(2)}(k, x, m) + r^{(2)} + r_E^{(2)}, \quad (7)$$

where $r_E^{(1)}$ and $r_E^{(2)}$ are environment induced noises which has mean zero but follow some temporal pattern rather than being random noise. The optimal leakage then becomes $(L^{(1)} - c^{(1)} - r_E^{(1)})(L^{(2)} - c^{(2)} - r_E^{(2)})$ instead. Therefore, we propose that the centering means $\bar{L}^{(1)}$ and $\bar{L}^{(2)}$ are calculated as moving averages from traces with a window of size n_w around the trace to be centered, rather than the average over all traces. The temporal patterns for $r_E^{(1)}$ and $r_E^{(2)}$, such as in Fig. 1, are usually slow changing. Hence, for a moderate window size, say $n_w = 100$, the moving averages $\bar{L}^{(1)} \approx c^{(1)} + r_E^{(1)}$ and $\bar{L}^{(2)} \approx c^{(2)} + r_E^{(2)}$.

When there are no environment induced noises $r_E^{(1)}$ and $r_E^{(2)}$, using bigger window size n_w can improve the precision. However, comparing to centering on averages of all traces, we can prove that centering the moving averages only loses $O(1/n_w)$ proportion of statistical efficiency under model (6). More precisely, denote the theoretical optimal leakage detection statistic as

$$\varDelta = (L_A^{(1)} - c^{(1)})(L_A^{(2)} - c^{(2)}) - (L_B^{(1)} - c^{(1)})(L_B^{(2)} - c^{(2)}). \qquad (8)$$

And denote the leakage detection statistic using moving average of a window size n_w as

$$D = (L_A^{(1)} - \bar{L}_A^{(1)})(L_A^{(2)} - \bar{L}_A^{(2)}) - (L_B^{(1)} - \bar{L}_B^{(1)})(L_B^{(2)} - \bar{L}_B^{(2)}). \qquad (9)$$

Then for large sample size n, the t-test statistic (3) is approximately $t_p(D) \approx E(D)/\sqrt{Var(D)/n}$, and the optimal leakage detection t-test statistic is approximately $t_p(\varDelta) \approx E(\varDelta)/\sqrt{Var(\varDelta)/n}$. A quantitative comparison of these two statistic is given in the next Theorem.

Theorem 1. *Under the second-order leakage model (6),*

$$\frac{E(D)}{\sqrt{Var(D)/n}} \frac{\sqrt{Var(\varDelta)/n}}{E(\varDelta)} = 1 - \frac{\eta}{n_w} + O(\frac{1}{n_w^2}), \qquad (10)$$

where the factor η is given by

$$\begin{aligned}\eta =& \frac{1}{Var(\varDelta)}[Var(V_A^{(1)})Var(V_A^{(2)}) + Var(V_B^{(1)})Var(V_B^{(2)}) + E^2(V_A^{(1)}V_A^{(2)}) \\ &+ E^2(V_B^{(1)}V_B^{(2)}) - Var(V_A^{(1)}V_A^{(2)}) - Var(V_B^{(1)}V_B^{(2)})].\end{aligned}$$

The proof of Theorem 1 is provided in Appendix A.

The factor η is usually small. When the noise variances σ_1^2 and σ_2^2 are big (so that the leakage is hard to detect), this factor $\eta = O[1/(\sigma_1^2\sigma_2^2)] \approx 0$. For practical situations, often $\eta < 1$. Hence using, say, $n_w = 100$ make the leakage detection statistic robust to environmental noises $r_E^{(1)}$ and $r_E^{(2)}$, at the price of a very small statistical efficiency loss when no environmental noises exist. Therefore, we recommend this paired moving-average based t-test (MA-t-test) over the existing tests.

We can also estimate the optimal window size n_w with some rough ideas of environmental noise fluctuation. The potential harm in using too wide a window is to introduce bias in the estimated centering quantities. Let the environmental noise be described as $r_E(t)$ for the $t = 1, 2, ..., T$ traces, and $\sum_{t=1}^{T} r_E(t) = 0$. Then the environmental noise induced bias in the moving average is bounded as $b \le a_0 n_w^2/2$ where a_0 is the maximum of the derivative $|r'_E(t)|$. Let $\varDelta_b^*$ denote the test statistic in Eq. (8) where the centering quantities $c^{(1)}$ and $c^{(2)}$ are each biased by the amount b. Then, (see Appendix B), $E(\varDelta_b^*) = E(\varDelta)$ and

$$\frac{Var(\varDelta_b^*)}{Var(\varDelta)} = 1 + \frac{b^2\eta^*}{Var(\varDelta)} + o(n_w^4) \le 1 + \frac{a_0^2 n_w^4 \eta^*}{4Var(\varDelta)} + o(n_w^4), \qquad (11)$$

bounds the harm of using a too big n_w value, where η^* is

$$Var(L_A^{(1)}) + Var(L_A^{(2)}) + Var(L_B^{(1)}) + Var(L_B^{(2)}) + 2E(V_A^{(1)}V_A^{(2)}) + 2E(V_B^{(1)}V_B^{(2)}).$$

Matching the Eqs. (10) and (11), we can estimate the optimal window size from $n_w^5 \approx$

$$\frac{4[Var(V_A^{(1)})Var(V_A^{(2)}) + Var(V_B^{(1)})Var(V_B^{(2)}) + E^2(V_A^{(1)}V_A^{(2)}) + E^2(V_B^{(1)}V_B^{(2)})]}{a_0^2[Var(L_A^{(1)}) + Var(L_A^{(2)}) + Var(L_B^{(1)}) + Var(L_B^{(2)}) + 2E(V_A^{(1)}V_A^{(2)}) + 2E(V_B^{(1)}V_B^{(2)})]}.$$

As an example, we estimate this window size using parameters for data sets reported in literature. For simplicity, we assume that both leakage time points follow a similar power model, $V_A^{(i)} = \epsilon[HW_i - E(HW)]$, $i = 1, 2$, with HW_i as hamming weights related to masks and plaintexts as in the model of [7,18]. Hence $E(V_A^{(1)}V_A^{(2)}) = 0$ can be dropped, and $Var(V_A^{(1)}) = \epsilon^2 Var(HW) = 2\epsilon^2$ for the one-byte hamming weight. With the signal-noise-ratio ϵ/σ around 0.1 as in [7,9], the noise variance dominates so that $Var(L_A^{(1)}) \approx \sigma^2$. Since the two groups A and B follows the same power model, the optimal window size formula is simplified to $\{4[2(2\epsilon^2)^2]/[a_0^2 4\sigma^2]\}^{1/5} = [8(\epsilon/\sigma)^4\sigma^2/a_0^2]^{1/5} = [8(0.1)^4\sigma^2/a_0^2]^{1/5}$. For the 2373-th time point on the traces on the DPA V2 contest data shown in Fig. 1, the environmental fluctuation is approximately four periods of sinusoidal curve over one million time points with magnitude ≈100. So taking the maximum derivative of this curve, $a_0 \approx 1/400$. Fitting the power model at this time point gets $\sigma \approx 300$. Hence the optimal window size here is $[400^2 8(0.1)^4 300^2]^{1/5} \approx 30$ traces. This optimal window size does vary with the magnitude of the environmental fluctuation and the leakage signal-noise-ratio which are not known to a tester as a prior. But this example can serve as a rough benchmark, and a window size of a few dozens may be used in practice.

3.3 Computational Efficiency

The paired t-test also has computational advantages over Welch's t-test. As pointed out in [19], computational stability can become an issue when using raw moments for large measurement campaigns. The paired t-test computes mean $\bar{D}$ and variance s_D^2 of local differences D. In case there is no detectable leakage, L_A and L_B have the same mean. Hence, the differences D are mean-free[1]. Even computing $\bar{D} = \frac{1}{n_i}\sum d_i$ is thus numerically stable. The sample variance s_D^2 can be computed as $s_D^2 = \overline{D^2} - (\bar{D})^2$, where only the first term $\overline{D^2}$ is not mean-free. We used the incremental equation from [17, Eq. (1.3)] to avoid numerical problems. Moreover, by applying the incremental equation for $\bar{D}$ as well, we were able to exploit straightforward parallelism when computing $\bar{D}$ and variance s_D^2.

[1] If D is not mean-free, a strong leakage exists. Hence, a small number of observations suffices for leakage detection, making numerical problems irrelevant.

Table 1. Computation accuracy between our incremental method and two-pass algorithm

	1st order	2nd order	3rd order	4th order	5th order
Our method	50.0097	2.4679e+3	4.5981e+5	7.3616e+7	1.7974e+10
Two pass	50.0097	2.4679e+3	4.5981e+5	7.3616e+7	1.7974e+10

The situation essentially remains the same for higher order or multivariate analysis: The differences D are still mean-free in the no-leakage case. Through the use of local averages, the three-pass approach is not necessary, since moving averages are used instead of global averages (cf. Eq. (9)). Computing moving averages is a local operation, as only nearby traces are considered. When processing traces in large blocks of e.g. 10k traces, all data needed for local averages is within the same file and can easily be accessed when needed, making the algorithm essentially one-pass. Similarly as in [19], we also give the experimental results using our method on 100 million simulated traces with $\sim\mathcal{N}(100, 25)$. Specifically, we compute the second parameters s_D^2 using the difference leakages: $D = L_A - L_B$ for first order test while $D = [(L_A - \bar{L}_{A,n_w})^d - (L_B - \bar{L}_{B,n_w})^d]$ for d-th order tests with moving average of window size $n_w = 100$. Table 1 shows our method matches the two-pass algorithm which computes the mean first and then the variance of the preprocessed traces. Note that D is not normalized using the central moment CM_2 and thus the second parameter is significantly larger than that in [19]. In the experiments, the same numerical stability is achieved without an extra pass, by focusing on the difference leakages.

4 Experimental Verification

To show the advantages of the new approach, the performances of the paired t-test (3) and the unpaired t-test (1) on synthetic data are compared.

First, we generate data for first order leakage according to model (4), where the environmental noise r_E follows a sinusoidal pattern similar to Fig. 1. The sinusoidal period is set as 200, 000 traces, and the sinusoidal magnitude is set as the pure measurement noise standard deviation $\sigma_A = \sigma_B = 50$. Hamming weight (HW) leakage is assumed in model (4). The first group A uses a fixed plaintext input corresponds to $HW = 5$, while the second group B uses random plaintexts. The paired t-test (3) and the unpaired t-test (1) are applied to the first $n = 30000, 60000, ..., 300000$ pairs of traces. The experiment is repeated 1000 times, and the proportions of leakage detection (rejection by each t-test) are plotted in Fig. 2.

Without any environmental noise r_E, the paired and unpaired t-tests perform the same. Their success rate curves overlap each other. With the sinusoidal noise r_E, the unpaired t-test uses many more traces to detect the leakage, while the paired t-test does not suffer from such performance degradation.

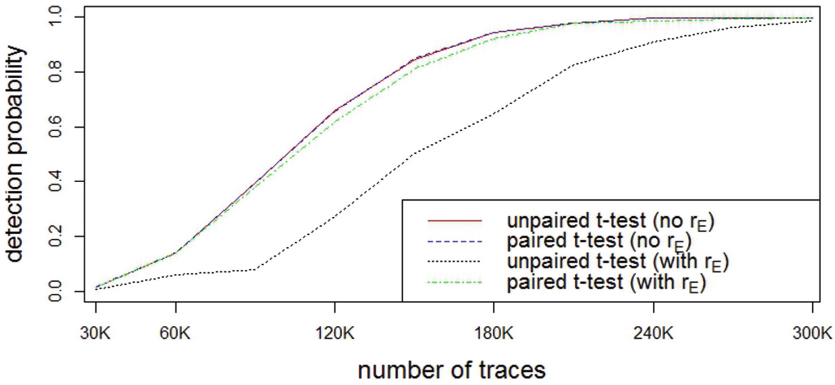


Fig. 2. T-test comparison for 1O leakage with and without a sinusoidal drift r_E. (Color figure online)

Notice that the environmental noise r_E often changes slowly as in Fig. 1. Hence, its effect is small for easy to detect leakage, when only a few hundreds or a few thousands of traces are needed. However, for hard to detect leakage, the effect has to be considered. We set a high noise level $\sigma_A = \sigma_B = 50$ to simulate a DUT with hard to detect first-order leakage. This allows the observable improvement by paired t-test over the unpaired t-test.

Second, we also generate data from the 2nd-order leakage model (7). The noise levels at the two leakage points, for both groups A and B, are set as $\sigma_1 = \sigma_2 = 10$ which are close to the levels in the physical implementation reported by [7]. We use the same sinusoidal environmental noise r_E as before. The first group A uses a fixed plaintext input corresponds to $HW = 1$, while the second group B uses random plaintexts. The proportions of leakage detection

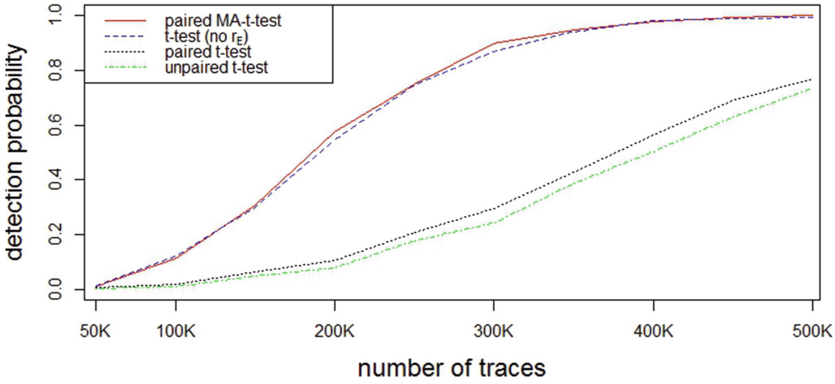


Fig. 3. T-test comparison for 2O leakage with a sinusoidal drift r_E. (Color figure online)

are plotted in Fig. 3. Again, we observe a serious degradation of t-test power to detect the leakage, when the environmental noise r_E is present. The paired t-test detects the leakage more often than the unpaired t-test in Fig. 3. However, the paired t-test also degrades comparing to the case without environmental noise r_E. That is due to the incorrect centering quantity for the 2O test as discussed in Sect. 3.2. Using the proposed method of centering at the moving average with window size 100, the paired MA-t-test has a performance close to the case where all environmental noise r_E is removed.

5 Practical Application

To show the advantage of the paired t-test in real measurement campaigns, we compare the performances of the unpaired and paired t-tests when analyzing an unprotected and an protected hardware implementation. The analysis focuses on the non-specific fixed vs. random t-test. We apply both tests to detect the first order leakage in the power traces acquired from an unprotected implementation of the NSA lightweight cipher Simon [2]. More specifically, a round-based implementation of Simon128/128 was used, which encrypts a 128-bit plaintext block with a 128-bit key in 68 rounds of operation. The second target is a masked engine of the same cipher. It is protected using three-share Threshold Implementation (TI) scheme, which is a round based variant of the TI Simon engine proposed in [20].

Both implementations are ported onto the SASEBO-GII board for power trace collection. The board is clocked at 3 MHz and a Tektronix oscilloscope samples the power consumption at 100 MS/s. Since Simon128/128 has 68 rounds, one power trace has about $68 \times \frac{1}{3\,\text{MHz}} \times 100\,\text{MS/s} \approx 2300$ time samples to cover the whole encryption and hence in the following experiments 2500 samples are taken in each measurement. The measurement setup is a modern setup that features a DC block and an amplifier. Note that the DC block will already take care of slow DC drifts that can affect the sensitivity of the unpaired t-test, as shown in Sect. 4. However, the DC block does not affect variations of the peak-to-peak height within traces, which are much more relevant for DPA. As the following experiments show, the paired t-test still shows improvement in such advanced setups.

5.1 Solving the Test-Order Bias

In [19], a random selection between fixed and random is proposed to avoid effects caused by states that occur in a fixed order, which we refer to as *test order*. For the paired (MA-)t-test, it is preferable to have a matching number of observations for both sets. We propose a fixed input sequence which is a repetition of $ABBA$ such that all the AB or BA pairs are constructed using neighboring inputs. For example in a sequence $ABBAABBA....ABBAABBA$, one alternately obtains AB and BA pairs with least variation. This ensures that all observations come in pairs and that the pairs are temporally close, so they share their environmental effects to a maximal possible degree. Moreover—even though the sequence is fixed and highly regular, the predecessor and successor for each measurement are perfectly balanced, corresponding to a 50 % probability of being either from the A or B set. This simpler setup removes the biases observed in [19] as efficiently as the random selection method. Experimental data of this section has been obtained using this scheme.

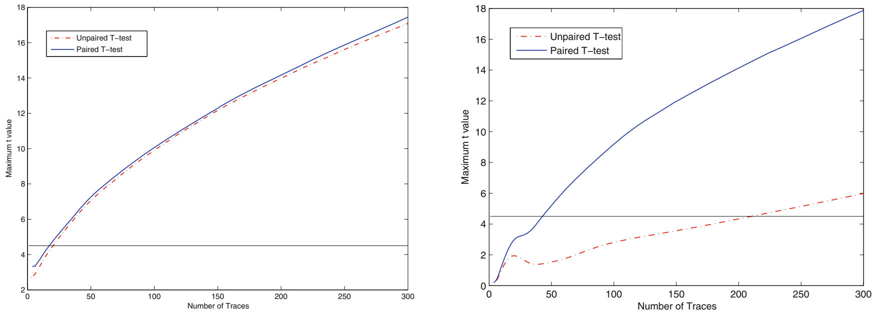
Note that the paired t-test can easily be applied in a random selection test order as well: After the trace collection, one can simply iteratively pair the leakages associated with the oldest fixed input and the oldest random input and then remove them from the sequence until no pairs can be constructed. An efficient way to do this is to separate all leakage traces into two subsets: $L_A = \{l_{A,1}, ...l_{A,n_A}\}$ and $L_B = \{l_{B,1}, ...l_{B,n_B}\}$ where $l_{A,i}$ and $l_{B,i}$ are the traces associated with i-th fixed input and i-th random input respectively in a chronological order and thus can be straightforwardly paired. Note that the cardinality of both sets are not always the same and hence only $n = min(n_A,\ n_B)$ AB pairs can be found. This approach is of less interest because time delay between fixed data and random data in a pair varies depending on the randomness of the input sequence.

5.2 First Order Analysis of an Unprotected Cipher

We first apply both paired and unpaired t-test to the unprotected engine which has strong first order leakage that can be exploited by DPA with only hundreds of traces. Usually the trace collection can be done quickly enough to avoid effects of environmental fluctuation in the measurements. However, to show the benefits of the paired t-test in this scenario, a hot air blower is used to heat up the crypto FPGA in SASEBO-GII board while the encryptions are executed. We designed two conditions to take the power measurements.

1. **Normal Lab Environment**, where measurements are performed in rapid succession, making the measurement campaign finish within seconds.
2. **Strong Environmental Fluctuation**, where a hot air blower was slowly moved towards and then away from the target FPGA to heat up and let it cool down again;

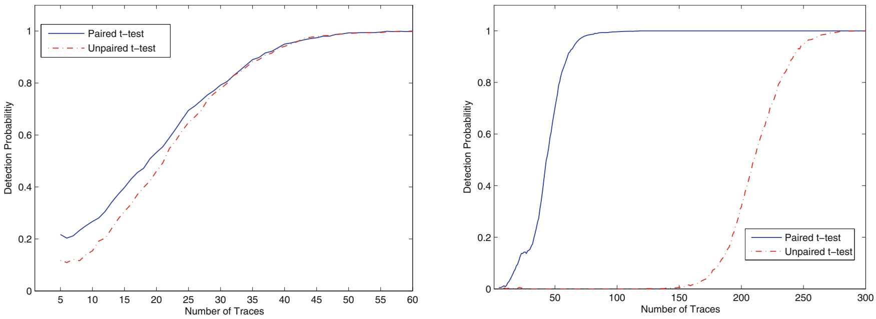
In each condition, 1000 measurements are taken alternately for the fixed plaintext and random plaintexts and later equally separated into two groups. In each group, the measurements are sorted in chronological order such that the j-th measurements of both groups are actually taken consecutively and share common variation. As explained in Sect. 5.1, the two measurements are a *matched-pair* and there are now 500 such pairs. Then both t-tests are applied to the first $n = 5, 6, 7, ..., 500$ pairs of measurements. For each n, the t-test returns a t-statistic vector of 2500 elements corresponding to 2500 time samples in the power traces because it is a univariate t-test. Our interest is the time sample that has the maximum t-statistic and thus the following results only focus on this specific time sample.



(a) No Environment Fluctuation (b) Environment Fluctuation

Fig. 4. T-test comparison for 1O leakage on unprotected Simon for a single measurement campaign of up to 300 pairs of traces. The paired t-test performs as well or better in both scenarios. However, the paired t-test is more robust to environmental noise. (Color figure online)

Figure 4 shows the t-statistics at the strongest leakage point as n increases. In Fig. 4(a) where there is no environmental fluctuation, both unpaired and paired t-test have the same performance as the t-statistic curves almost overlap. However, in Fig. 4(b) where the varying temperature changed the power traces greatly, the paired t-test (blue solid line) shows robustness and requires less traces to exceed the threshold of 4.5 while the performance of the unpaired t-test is greatly reduced in the sense that more traces are needed to go beyond the threshold. Figure 5 shows the detection probability of the t-tests in the same scenario. First, 1000 repetitions of the above experiment are performed and the number of experiments that result in a t-statistic above 4.5 is counted. Detection probability equals this number divided by 1000. Figure 5(a) shows the detection



(a) No Environment Fluctuation (b) Environment Fluctuation

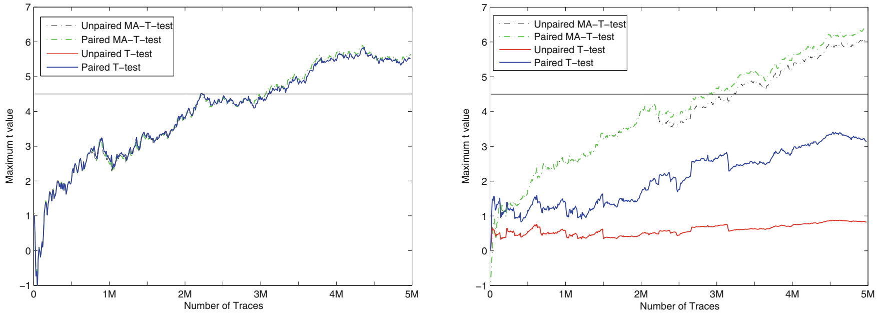
Fig. 5. T-test detection probability for 10 leakage. Again, the paired t-test performs at least as well as the unpaired, while being much more robust in the presence of environmental noise.

probability of two tests under normal lab condition. With more than 30 pairs, both tests can detect the first order leakage with the same probability. With more than 60 pairs the detection probability rises to 1 for both tests which shows the efficiency of both tests on the normal traces. Figure 5(b) shows that paired t-test (solid line) is still robust in spite of varying environmental factors. With less than 100 pairs, the detection probability of paired t-test is already 1 while unpaired t-test requires much more traces to achieve the same probability.

In summary, the paired t-test is more robust and efficient in detecting first order leakage when the power traces are collected in a quickly changing environment.

5.3 Second Order Analysis on a First-Order Resistant Design

In order to validate the effectiveness of the paired t-test in a longer measurement campaign, where environmental fluctuations are very likely to occur, a first-order-leakage-resistant Simon engine protected by a three-share Threshold Implementation scheme is used as the target. Five million power traces are collected in a room without windows and without expected fluctuations in temperature over a period 5 h. As before, one measurement campaign is performed in a stable lab environment where the environmental conditions are kept as stable as possible. In the other scenario, we again used the hot air blower in intervals of several minutes to simulate stronger environmental noise. This is because the environmental noise might not be strong during the 5-h collection period. However, in scenarios where hundreds of millions of measurements are needed and taken over a period of several days, then environmental fluctuation can be found, as in Fig. 1.



(a) No Environment Fluctuation

(b) Environment Fluctuation

Fig. 6. T-test detection probability for 20 leakage (Color figure online)

As before, the 5 million traces are equally divided into two groups for fixed and random plaintext respectively. The first order t-test does not indicate any leakage ($|t| < 3$), as expected. Figure 6 shows the t-statistics of the second order t-tests as the number of traces increases in both the stable lab environment and the simulated lab environment noise scenario. In the first experiment in a stable environment, depicted in Fig. 6(a), we compare both tests using global average and moving average. The curve of four tests almost overlap and they perform about the same with about three million traces needed to achieve a t-statistic above 4.5. This shows that paired t-test works as well as unpaired one for constant collection environment. Also, the moving average based tests perform very similar to the global average based tests, with a minor improvement in the relevant many-traces case. Figure 6(b) depicts the results for the experiment with strong environmental fluctuations. The paired MA-t-test performs best and goes beyond 4.5 faster than the unpaired one. The other two tests using global average are still below the threshold with 5 million traces. The paired t-test still clearly outperforms the unpaired t-test. In sum, the paired t-test based on moving average is the most robust to fluctuation and significantly improves the performance of higher order analysis.

6 Conclusion

Welch's t-test has recently received a lot of attention as standard side channel security evaluation tool. In this work we showed that noise resulting from environmental fluctuations can negatively impact the performance of Welch's t-test. The resulting increased number of observations to detect a leakage are inconvenient and can, in the worst case, result in false conclusions about a device's resistance. We proposed a paired t-test to improve the standard methodology for leakage detection. The resulting matched-pairs design removes the environmental noise effect in leakage detection. Furthermore, we showed that moving averages increase the robustness against environmental noise for higher order or multivariate analysis, while not showing any negative impact in the absence of noise. The improvement is shown through mathematical analysis, simulation, and on practical power measurements: both paired and unpaired t-test with and without the moving averages approach are compared for first order and second order analysis. Our results show that the proposed (moving average based) paired t-test performed as well or better in all analyzed scenarios. The new method does not increase computational complexity and is numerically more stable than Welch's t-test. Since our method is more robust to environmental noise and can detect leakage faster than unpaired test in the presence of noise, we propose the replacement of Welch's t-test with the moving average based paired t-test as a standard leakage detection tool.

Acknowledgments. This work is supported by the National Science Foundation under grant CNS-1314655, CNS-1314770 and CNS-1261399.

Appendix

A Proof of Theorem 1

We are comparing the leakage detection statistic (9)

$$D = (L_A^{(1)} - \bar{L}_A^{(1)})(L_A^{(2)} - \bar{L}_A^{(2)}) - (L_B^{(1)} - \bar{L}_B^{(1)})(L_B^{(2)} - \bar{L}_B^{(2)}),$$

with the theoretical optimal leakage detection statistic Δ in Eq. (8).

Without loss of generality, let $c^{(1)} = c^{(2)} = 0$ in model (6), since these constants are cancelled in each of the differences $(L_A^{(j)} - \bar{L}_A^{(j)})$ and $(L_B^{(j)} - \bar{L}_B^{(j)})$ for $j = 1, 2$. Then (8) is simplified as $\Delta = L_A^{(1)} L_A^{(2)} - L_B^{(1)} L_B^{(2)}$. Hence

$$\begin{aligned} E(\Delta) &= E(L_A^{(1)} L_A^{(2)}) - E(L_B^{(1)} L_B^{(2)}) \\ Var(\Delta) &= Var(L_A^{(1)} L_A^{(2)}) + Var(L_B^{(1)} L_B^{(2)}). \end{aligned} \tag{12}$$

We first reexpress $(L_A^{(1)} - \bar{L}_A^{(1)})$ as the difference between two independent terms. We denote $\tilde{L}_A^{(1)} = \frac{1}{n_w - 1} \sum_{i=1}^{n_w - 1} L_{A,i}^{(1)}$ as the average of $n_w - 1$ traces excluding the original trace, where $L_{A,i}^{(1)}$ $(i = 1, ..., n_w - 1)$ are independent random variables coming from the same distribution as $L_A^{(1)}$. Since $\bar{L}_A^{(1)}$ is the average over n_w nearby traces including the original trace, $\bar{L}_A^{(1)} = \frac{1}{n_w}[L_A^{(1)} + \sum_{i=1}^{n_w-1} L_{A,i}^{(1)}] = \frac{n_w - 1}{n_w}(L_A^{(1)} - \tilde{L}_A^{(1)})$, with $\tilde{L}_A^{(1)}$ independent of $L_A^{(1)}$. $E(\tilde{L}_A^{(1)}) = E(L_A^{(1)})$ and $Var(\tilde{L}_A^{(1)}) = \frac{1}{n_w - 1} Var(L_A^{(1)})$. Similarly, $\tilde{L}_A^{(2)}$, $\tilde{L}_B^{(1)}$ and $\tilde{L}_B^{(2)}$ denotes the average of corresponding quantities over the $n_w - 1$ traces excluding the original trace. The we can rewrite the leakage detection statistic in (9) as

$$D = (\frac{n_w - 1}{n_w})^2 [(L_A^{(1)} - \tilde{L}_A^{(1)})(L_A^{(2)} - \tilde{L}_A^{(2)}) - (L_B^{(1)} - \tilde{L}_B^{(1)})(L_B^{(2)} - \tilde{L}_B^{(2)})]. \tag{13}$$

Therefore as $n_w \to \infty$, $D \to \Delta$.

Next, we show that $E(D)$ and $Var(D)$ differ from their limits $E(\Delta)$ and $Var(\Delta)$ by a factor of $O(1/n_w)$ only. Let $D^* = \frac{n_w}{n_w - 1} D$. Then we have

$$E(D^*) = E(\Delta), \tag{14}$$

$$\begin{aligned} &Var(D^*) - Var(\Delta) \\ =&\frac{2}{n_w}[Var(V_A^{(1)})Var(V_A^{(2)}) + Var(V_B^{(1)})Var(V_B^{(2)}) + E^2(V_A^{(1)} V_A^{(2)}) \\ &+ E^2(V_B^{(1)} V_B^{(2)}) - Var(V_A^{(1)} V_A^{(2)}) - Var(V_B^{(1)} V_B^{(2)})] + O(\frac{1}{n_w^2}). \end{aligned} \tag{15}$$

The proofs of these two equations are provided in the next two subsections.

Combining Eqs. (12), (14) and (15), we arrived at Eq. (10) and Theorem 1 is proved.

A.1 Proof of Eq. (14) on Mean of D^*

We now calculate the first term in $E(D)$.

$$E(\tilde{L}_A^{(1)}\tilde{L}_A^{(2)}) = (\frac{1}{n_w-1})^2 \sum_{i=1}^{n_w-1}\sum_{j=1}^{n_w-1} E(L_{A,i}^{(1)}L_{A,j}^{(2)}).$$

For $i \neq j$, $L_{A,i}^{(1)}$ is independence of $L_{A,j}^{(2)}$ so that $E(L_{A,i}^{(1)}L_{A,j}^{(2)}) = E(L_{A,i}^{(1)})E(L_{A,j}^{(2)}) = (0)(0) = 0$ and drops from the summation. Hence

$$E(\tilde{L}_A^{(1)}\tilde{L}_A^{(2)}) = (\frac{1}{n_w-1})^2 \sum_{i=1}^{n_w-1} E(L_{A,i}^{(1)}L_{A,i}^{(2)}) = \frac{1}{n_w-1}E(L_A^{(1)}L_A^{(2)}). \tag{16}$$

Also, since $\tilde{L}_A^{(1)}$ is independent of $L_A^{(2)}$, $E(\tilde{L}_A^{(1)}L_A^{(2)}) = E(\tilde{L}_A^{(1)})E(L_A^{(2)}) = 0$. Similarly $E(L_A^{(1)}\tilde{L}_A^{(2)}) = 0$. Therefore,

$$\begin{aligned} E[(L_A^{(1)} - \tilde{L}_A^{(1)})(L_A^{(2)} - \tilde{L}_A^{(2)})] &= E(L_A^{(1)}L_A^{(2)}) - 0 - 0 + E(\tilde{L}_A^{(1)}\tilde{L}_A^{(2)}) \\ &= E(L_A^{(1)}L_A^{(2)}) + \frac{1}{n_w-1}E(L_A^{(1)}L_A^{(2)}) \\ &= \frac{n_w}{n_w-1}E(L_A^{(1)}L_A^{(2)}). \end{aligned}$$

Similarly, $E[(L_B^{(1)} - \tilde{L}_B^{(1)})(L_B^{(2)} - \tilde{L}_B^{(2)})] = \frac{n_w}{n_w-1}E(L_B^{(1)}L_B^{(2)})$. Combine these two expressions with Eq. (13) and $D^* = \frac{n_w}{n_w-1}D$, we get Eq. (14)

$$E(D^*) = (\frac{n_w-1}{n_w})\frac{n_w}{n_w-1}E[L_A^{(1)}L_A^{(2)} - L_B^{(1)}L_B^{(2)}] = E(\Delta).$$

A.2 Proof of Eq. (15) on Variance of D^*

$$Var(D^*) = (\frac{n_w-1}{n_w})^2\{Var[(L_A^{(1)}-\tilde{L}_A^{(1)})(L_A^{(2)}-\tilde{L}_A^{(2)})]+Var[(L_B^{(1)}-\tilde{L}_B^{(1)})(L_B^{(2)}-\tilde{L}_B^{(2)})]\}. \tag{17}$$

For the first term, the variance of the sum $L_A^{(1)}L_A^{(2)} - \tilde{L}_A^{(1)}L_A^{(2)} - L_A^{(1)}\tilde{L}_A^{(2)} + L_A^{(1)}L_A^{(2)}$ is the covariance of the sum with itself. For the four terms in $L_A^{(1)}L_A^{(2)} - \tilde{L}_A^{(1)}L_A^{(2)} - L_A^{(1)}\tilde{L}_A^{(2)} + L_A^{(1)}L_A^{(2)}$, the covariance for most pairs of different terms are zero. For example,

$$\begin{aligned} Cov(L_A^{(1)}L_A^{(2)}, \tilde{L}_A^{(1)}L_A^{(2)}) &= E(L_A^{(1)}L_A^{(2)}\tilde{L}_A^{(1)}L_A^{(2)}) - E(L_A^{(1)}L_A^{(2)})E(\tilde{L}_A^{(1)}L_A^{(2)}) \\ &= E(L_A^{(1)}L_A^{(2)}L_A^{(2)})0 - E(L_A^{(1)}L_A^{(2)})E(L_A^{(2)})0 = 0. \end{aligned}$$

and $Cov(L_A^{(1)}L_A^{(2)}, \tilde{L}_A^{(1)}\tilde{L}_A^{(2)}) = 0$ due to the independence between $L_A^{(1)}L_A^{(2)}$ and $\tilde{L}_A^{(1)}\tilde{L}_A^{(2)}$. The only non-zero cross-term covariance is

$$Cov(\tilde{L}_A^{(1)}L_A^{(2)}, L_A^{(1)}\tilde{L}_A^{(2)}) = E(\tilde{L}_A^{(1)}L_A^{(2)}L_A^{(1)}\tilde{L}_A^{(2)}) - 0 = E(L_A^{(1)}L_A^{(2)})E(\tilde{L}_A^{(1)}\tilde{L}_A^{(2)})$$
$$= \frac{1}{n_w - 1}E^2(L_A^{(1)}L_A^{(2)}),$$

with the last step coming from Eq. (16). Therefore,

$$\begin{aligned} &Var[(L_A^{(1)} - \tilde{L}_A^{(1)})(L_A^{(2)} - \tilde{L}_A^{(2)})] \\ =&Var(L_A^{(1)}L_A^{(2)}) + Var(\tilde{L}_A^{(1)}L_A^{(2)}) + Var(L_A^{(1)}\tilde{L}_A^{(2)}) + Var(\tilde{L}_A^{(1)}\tilde{L}_A^{(2)}) \\ &+ \frac{2}{n_w - 1}E^2(L_A^{(1)}L_A^{(2)}) \end{aligned}$$

By independence, $Var(\tilde{L}_A^{(1)}L_A^{(2)}) = Var(\tilde{L}_A^{(1)})Var(L_A^{(2)}) = \frac{1}{n_w-1}Var(L_A^{(1)})$ $Var(L_A^{(2)})$, and $Var(L_A^{(1)}\tilde{L}_A^{(2)}) = \frac{1}{n_w-1}Var(L_A^{(1)})Var(L_A^{(2)})$.

For $Var(\tilde{L}_A^{(1)}\tilde{L}_A^{(2)})$, note that

$$\tilde{L}_A^{(1)}\tilde{L}_A^{(2)} = (\frac{1}{n_w - 1})^2 \sum_{i=1}^{n_w-1}\sum_{j=1}^{n_w-1} L_{A,i}^{(1)}L_{A,j}^{(2)}.$$

The covariance between any two different terms in the sum is zero. Hence

$$\begin{aligned} Var(\tilde{L}_A^{(1)}\tilde{L}_A^{(2)}) &= (\frac{1}{n_w - 1})^4[\sum_i Var(L_{A,i}^{(1)}L_{A,i}^{(2)}) + \sum_{i\neq j} Var(L_{A,i}^{(1)}L_{A,j}^{(2)})] \\ &= \frac{1}{(n_w - 1)^3}Var(L_A^{(1)}L_A^{(2)}) + \frac{n_w - 2}{(n_w - 1)^3}Var(L_A^{(1)})Var(L_A^{(2)}). \end{aligned}$$

Combine together, we have

$$\begin{aligned} &Var[(L_A^{(1)} - \tilde{L}_A^{(1)})(L_A^{(2)} - \tilde{L}_A^{(2)})] \\ &= Var(L_A^{(1)}L_A^{(2)}) + \frac{2}{n_w - 1}Var(L_A^{(1)})Var(L_A^{(2)}) + \frac{2}{n_w - 1}E^2(L_A^{(1)}L_A^{(2)}) \\ &\quad + \frac{n_w - 2}{(n_w - 1)^3}Var(L_A^{(1)})Var(L_A^{(2)}) + \frac{1}{(n_w - 1)^3}Var(L_A^{(1)}L_A^{(2)}) \\ &= Var(L_A^{(1)}L_A^{(2)}) + \frac{2}{n_w}Var(L_A^{(1)})Var(L_A^{(2)}) + \frac{2}{n_w}E^2(L_A^{(1)}L_A^{(2)}) + O(\frac{1}{n_w^2}) \end{aligned}$$

Hence the first term in $Var(D^*)$ becomes

$$\begin{aligned} &(\frac{n_w - 1}{n_w})^2Var[(L_A^{(1)} - \tilde{L}_A^{(1)})(L_A^{(2)} - \tilde{L}_A^{(2)})] \\ =&(\frac{n_w - 1}{n_w})^2Var(L_A^{(1)}L_A^{(2)}) + \frac{2}{n_w}Var(L_A^{(1)})Var(L_A^{(2)}) + \frac{2}{n_w}E^2(L_A^{(1)}L_A^{(2)}) + O(\frac{1}{n_w^2}) \\ =&Var(L_A^{(1)}L_A^{(2)}) + \frac{2}{n_w}[Var(L_A^{(1)})Var(L_A^{(2)}) + E^2(L_A^{(1)}L_A^{(2)}) - Var(L_A^{(1)}L_A^{(2)})] + O(\frac{1}{n_w^2}). \end{aligned} \tag{18}$$

For further simplification, let σ_1^2 and σ_2^2 denote the variances of noises $r^{(1)}$ and $r^{(2)}$ in the second-order leakage model (6). Then $Var(L_A^{(1)}) = \sigma_1^2 + Var(V^{(1)})$, $Var(L_A^{(2)}) = \sigma_2^2 + Var(V^{(2)})$, $E(L_A^{(1)}L_A^{(2)}) = E(V^{(1)}V^{(2)})$,

$$\begin{aligned}&E[(L_A^{(1)}L_A^{(2)})^2] = E[(V_A^{(1)} + r_A^{(1)})^2(V_A^{(2)} + r_A^{(2)})^2]\\ =&E[(V_A^{(1)})^2(V_A^{(2)})^2 + (r_A^{(1)})^2(V_A^{(2)})^2 + (V_A^{(1)})^2(r_A^{(2)})^2 + (r_A^{(1)})^2(r_A^{(2)})^2] + 0\\ =&E[(V_A^{(1)})^2(V_A^{(2)})^2] + \sigma_1^2 Var(V_A^{(2)}) + \sigma_2^2 Var(V_A^{(1)}) + \sigma_1^2\sigma_2^2.\end{aligned}$$

Hence

$$Var[L_A^{(1)}L_A^{(2)}] = Var(V_A^{(1)}V_A^{(2)}) + \sigma_1^2 Var(V_A^{(2)}) + \sigma_2^2 Var(V_A^{(1)}) + \sigma_1^2\sigma_2^2.$$

Combine the above five expressions,

$$\begin{aligned}&Var(L_A^{(1)})Var(L_A^{(2)}) + E^2(L_A^{(1)}L_A^{(2)}) - Var(L_A^{(1)}L_A^{(2)})\\ =&Var(V^{(1)})Var(V^{(2)}) + E(V^{(1)}V^{(2)}) - Var(V_A^{(1)}V_A^{(2)})\end{aligned}$$

Combine this with (17) and (18) we have Eq. (15),

$$\begin{aligned}&Var(D^*) - [Var(L_A^{(1)}L_A^{(2)}) + Var(L_B^{(1)}L_B^{(2)})]\\ &= \frac{2}{n_w}[Var(V_A^{(1)})Var(V_A^{(2)}) + E^2(V_A^{(1)}V_A^{(2)}) - Var(V_A^{(1)}V_A^{(2)})\\ &\quad + Var(V_B^{(1)})Var(V_B^{(2)}) + E^2(V_B^{(1)}V_B^{(2)}) - Var(V_B^{(1)}V_B^{(2)})] + O(\frac{1}{n_w^2}).\end{aligned}$$

B Derivation of Eq. (11)

As in the previous section, we let $c^{(1)} = c^{(2)} = 0$ without loss of generality, so that $E(L_A^{(1)}) = E(L_A^{(2)}) = 0$. Then

$$\begin{aligned}E[(L_A^{(1)} - b)(L_A^{(2)} - b)] &= E(L_A^{(1)}L_A^{(2)}) - bE(L_A^{(1)}) - bE(L_A^{(2)}) + b^2 = E(L_A^{(1)}L_A^{(2)}) + b^2\\ &= E(L_A^{(1)}L_A^{(2)}) + b^2.\end{aligned}$$

Hence

$$\begin{aligned}E(\varDelta_b^*) &= E[(L_A^{(1)} - b)(L_A^{(2)} - b)] - E[(L_B^{(1)} - b)(L_B^{(2)} - b)]\\ &= E(L_A^{(1)}L_A^{(2)}) + b^2 - E(L_B^{(1)}L_B^{(2)}) - b^2\\ &= E(L_A^{(1)}L_A^{(2)}) - E(L_B^{(1)}L_B^{(2)}) = E(\varDelta). \qquad (19)\end{aligned}$$

Next,

$$
\begin{aligned}
&Var[(L_A^{(1)}-b)(L_A^{(2)}-b)]\\
=&E[(L_A^{(1)}-b)^2(L_A^{(2)}-b)^2]-[E(L_A^{(1)}L_A^{(2)})+b^2]^2\\
=&E[((L_A^{(1)})^2-2bL_A^{(1)}+b^2)((L_A^{(2)})^2-2bL_A^{2)}+b^2)]-E[(L_A^{(1)}L_A^{(2)})^2]-2bE(L_A^{(1)}L_A^{(2)})-b^4\\
=&Var(L_A^{(1)}L_A^{(2)})-2bE[L_A^{(1)}L_A^{(2)}(L_A^{(1)}+L_A^{(2)})]+b^2E[(L_A^{(1)})^2+(L_A^{(2)})^2+2L_A^{(1)}L_A^{(2)})]\\
=&Var(L_A^{(1)}L_A^{(2)})+b^2[Var(L_A^{(1)})+Var(L_A^{(2)})+2E(L_A^{(1)}L_A^{(2)})]+O(b).
\end{aligned}
$$

Hence we get the variance

$$
\begin{aligned}
Var(\Delta_b^*)=&Var(\Delta)+b^2[Var(L_A^{(1)})+Var(L_A^{(2)})+2E(L_A^{(1)}L_A^{(2)})\\
&+Var(L_B^{(1)})+Var(L_B^{(2)})+2E(L_B^{(1)}L_B^{(2)})]+O(b).
\end{aligned}
\tag{20}
$$

References

1. Balasch, J., Gierlichs, B., Grosso, V., Reparaz, O., Standaert, F.-X.: On the cost of lazy engineering for masked software implementations. In: Joye, M., Moradi, A. (eds.) CARDIS 2014. LNCS, vol. 8968, pp. 64–81. Springer, Heidelberg (2015). http://dx.doi.org/10.1007/978-3-319-16763-3_5
2. Beaulieu, R., Shors, D., Smith, J., Treatman-Clark, S., Weeks, B., Wingers, L.: The Simon and Speck families of lightweight block ciphers. IACR Cryptol. ePrint Arch. **2013**, 404 (2013)
3. Bilgin, B., Gierlichs, B., Nikova, S., Nikov, V., Rijmen, V.: A more efficient AES threshold implementation. In: Pointcheval, D., Vergnaud, D. (eds.) AFRICACRYPT. LNCS, vol. 8469, pp. 267–284. Springer, Heidelberg (2014)
4. Bilgin, B., Gierlichs, B., Nikova, S., Nikov, V., Rijmen, V.: Higher-order threshold implementations. In: Sarkar, P., Iwata, T. (eds.) ASIACRYPT 2014, Part II. LNCS, vol. 8874, pp. 326–343. Springer, Heidelberg (2014)
5. Chen, C., Eisenbarth, T., von Maurich, I., Steinwandt, R.: Masking large keys in hardware: a masked implementation of McEliece. In: Dunkelman, O., et al. (eds.) SAC 2015. LNCS, vol. 9566, pp. 293–309. Springer, Heidelberg (2016). doi:10.1007/978-3-319-31301-6_18
6. Cooper, J., DeMulder, E., Goodwill, G., Jaffe, J., Kenworthy, G., Rohatgi, P.: Test Vector Leakage Assessment (TVLA) methodology in practice. In: International Cryptographic Module Conference (2013). http://icmc-2013.org/wp/wp-content/uploads/2013/09/goodwillkenworthtestvector.pdf
7. Ding, A.A., Zhang, L., Fei, Y., Luo, P.: A statistical model for higher order DPA on masked devices. In: Batina, L., Robshaw, M. (eds.) CHES 2014. LNCS, vol. 8731, pp. 147–169. Springer, Heidelberg (2014). http://dx.doi.org/10.1007/978-3-662-44709-3_9
8. Durvaux, F., Standaert, F.-X.: From improved leakage detection to the detection of points of interests in leakage traces. In: Fischlin, M., Coron, J.-S. (eds.) EUROCRYPT 2016. LNCS, vol. 9665, pp. 240–262. Springer, Heidelberg (2016). doi:10.1007/978-3-662-49890-3_10
9. Fei, Y., Ding, A.A., Lao, J., Zhang, L.: A statistics-based success rate model for DPA and CPA. J. Crypt. Eng. **5**(4), 227–243 (2015). doi:10.1007/s13389-015-0107-0

10. Goodwill, G., Jun, B., Jaffe, J., Rohatgi, P.: A testing methodology for side-channel resistance validation. In: NIST Non-Invasive Attack Testing Workshop, September 2011. http://csrc.nist.gov/news_events/non-invasive-attack-testing-workshop/papers/08_Goodwill.pdf
11. Heuser, A., Kasper, M., Schindler, W., Stöttinger, M.: A new difference method for side-channel analysis with high-dimensional leakage models. In: Dunkelman, O. (ed.) CT-RSA 2012. LNCS, vol. 7178, pp. 365–382. Springer, Heidelberg (2012). http://dx.doi.org/10.1007/978-3-642-27954-6_23
12. Kutner, M.H., Nachtsheim, C.J., Neter, J., Li, W.: Applied Linear Statistical Models. McGraw-Hill/Irwin, New York (2005)
13. Leiserson, A.J., Marson, M.E., Wachs, M.A.: Gate-level masking under a path-based leakage metric. In: Batina, L., Robshaw, M. (eds.) CHES 2014. LNCS, vol. 8731, pp. 580–597. Springer, Heidelberg (2014)
14. Mather, L., Oswald, E., Bandenburg, J., Wójcik, M.: Does my device leak information? an *a priori* statistical power analysis of leakage detection tests. In: Sako, K., Sarkar, P. (eds.) ASIACRYPT 2013, Part I. LNCS, vol. 8269, pp. 486–505. Springer, Heidelberg (2013). http://dx.doi.org/10.1007/978-3-642-42033-7_25
15. Moradi, A., Hinterwälder, G.: Side-channel security analysis of ultra-low-power FRAM-based MCUs. In: Mangard, S., Poschmann, A.Y. (eds.) COSADE 2015. LNCS, vol. 9064, pp. 239–254. Springer, Heidelberg (2015). http://dx.doi.org/10.1007/978-3-319-21476-4_16
16. Nascimento, E., Lopez, J., Dahab, R.: Efficient and secure elliptic curve cryptography for 8-bit AVR microcontrollers. In: Chakraborty, R.S., et al. (eds.) SPACE 2015. LNCS, vol. 9354. Springer, Heidelberg (2015). http://dx.doi.org/10.1007/978-3-319-24126-5_17
17. Pébay, P.: Formulas for robust, one-pass parallel computation of covariances and arbitrary-order statistical moments. Sandia report SAND2008-6212, Sandia National Laboratories (2008)
18. Prouff, E., Rivain, M., Bevan, R.: Statistical analysis of second order differential power analysis. IEEE Trans. Comput. **58**(6), 799–811 (2009)
19. Schneider, T., Moradi, A.: Leakage assessment methodology. In: Güneysu, T., Handschuh, H. (eds.) CHES 2015. LNCS, vol. 9293, pp. 495–513. Springer, Heidelberg (2015). http://dblp.uni-trier.de/db/conf/ches/ches2015.html SchneiderM15
20. Shahverdi, A., Taha, M., Eisenbarth, T.: Silent Simon: threshold implementation under 100 slices. In: 2015 IEEE International Symposium on Hardware Oriented Security and Trust (HOST), pp. 1–6, May 2015