THE GOOD, THE BAD, ALGORITHMIC NOISE TOLERANCE (ANT), THE UGLY

Noyan C. Sevüktekin and Andrew C. Singer

University of Illinois at Urbana-Champaign
Coordinated Science Laboratory, Department of Electrical and Computer Engineering

ABSTRACT

Computational units implemented on nanoscale physical substrates are susceptible to errors that can be catastrophic if not mitigated. Statistical error compensation techniques have become prevalent to safeguard computational units against such hardware-failures. Algorithmic Noise Tolerance (ANT) is one such technique that utilizes a low-fidelity replica unit to detect and bypass such failures occurring within the primary (main) computational unit. Connections between ANT and the binary hypothesis testing as well as the information theoretic CEO problem have been explored for sub-exponential error profiles, quadratic and logarithmic distortion functions. However, there exist fundamental performance limits of ANT approach even without such model-dependent restrictions. The purpose of this paper is to explore fidelity-dependent conditions that are universal over the statistical properties of the computational units under which, the overall performance of ANT is arbitrarily close to the fundamental limits.

Index Terms— Algorithmic Noise Tolerance, Statistical Error Compensation, Calibration, Decision, Mixture Models.

1. INTRODUCTION

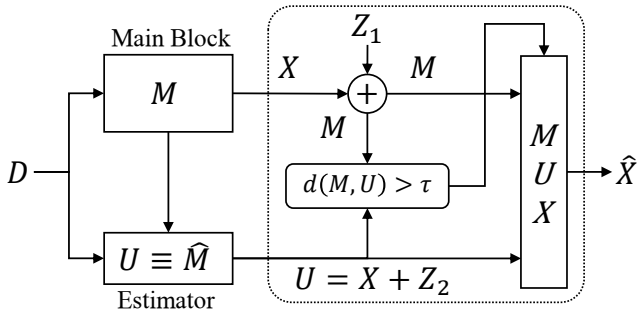
The trade-off between precision and reliability in nanoscale systems is a prominent phenomenon challenging the exponential improvement driven by the Moore's law. As the process technology feature sizes scale down, the computational uncertainties due to process, temperature and voltage variations create reliability bottlenecks [1]. Statistical error compensation (SEC) techniques motivated by information and decision-theoretic frameworks provide robust, low-power solutions at the subsystem level as an alternative to modular redundancy-based error compensation techniques, which are often power-hungry [2].

Algorithmic noise tolerance (ANT) is an SEC technique that safeguards a high-precision computational unit that is prone to hardware failures by detecting and bypassing such errors via a low-precision yet robust calibration unit. ANT builds decision statistics at low-power by measuring the distance between the outcome of the main computational unit and that of the calibration unit, and it bypasses the erroneous outcomes by switching between the units. ANT has been shown to reduce the overall power consumption by over 50% at the expense of logic overhead of less than 10%, while maintaining false alarm probability, $\pi_{1|0} < 0.1$ and miss probability, $\pi_{0|1} < 0.01$, [3–5]. These practical improvements motivate the search for the fundamental limits of ANT as well as a generalization of the replica-block based approach for low-power statistical decision making.

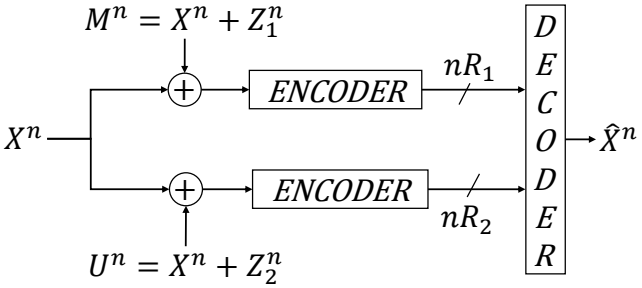
A decision theoretic analysis of ANT with Bayesian priors has been given in [3, 6]. An information theoretic analysis establishing a connection between ANT and the Gaussian CEO problem was given in [7]. As discussed in Section 2, these analyses provide characterizations of specific error statistics that can be eliminated using the ANT architecture. However, even without statistical restrictions the following question can be addressed: *How can the binary hypotheses be tested while only having access to external statistics?* The purpose of this paper is to explore fidelity-based criteria for the regions of operation for a generalized ANT model, where the performance of an *oracle* that has access to hardware failure information is arbitrarily close to that of ANT.

This paper addresses the following fundamental challenges: The current analyzed model of ANT is limited to an additive noise model and does not generalize to arbitrary hypothesis testing frameworks and fidelity-based performance criteria, such as those discussed in [8], for ANT are unknown. In the typical analysis model shown in Fig. 1 (a), ANT comprises a main unit followed by hardware errors and an estimator unit, prone to estimation errors. Instead, we propose a mixture model for the main block, seen in Fig. 2, mixing a *good* computational unit with a *bad* one and instead of an estimator of the main block, we model the calibration unit as a robust yet lower-fidelity unit called the *ugly* unit. We further propose criteria for the optimality and near optimality of the ANT architecture via a fidelity-based ordering of each unit and make the connection between the expected loss of each unit and the expected distance between their outputs.

This paper is organized as follows: Section 2 discusses the relation to previous work in detail. Section 3 introduces our model and provides the mathematical foundations of our approach. Section 4 discussed fidelity-based performance limits of a notional ANT architecture. Section 5 outlines proofs. A set of numerical examples are given in Section 6.



(a) Main-estimator block model for ANT architecture [3, 6]



(b) Analogous model in the Gaussian CEO framework [7].

Fig. 1: Additive noise models for the ANT architecture from decision and information theoretic frameworks

2. RELATION TO PRIOR WORK

The idea of low-power SEC goes back to [2, 9], where soft-error cancellation was proposed for digital signal processing (DSP) and low-power filtering. A decision theoretic analysis of ANT was given in [6] and later improved upon in [3], under the model given in Fig. 1 (a), [6, Fig. 1-2], for which error statistics were assumed to belong to a sub-exponential family, yielding that the distance between outcomes provides the sufficient statistics [10]. An information theoretic exploration of certain fundamental limits of ANT was given in [7] by making an analogy between the ANT setup and the Gaussian CEO problem as seen in Fig. 1 (b), [7, Fig. 1], under quadratic and logarithmic distortion. Both of these models propose additive hardware and estimation errors. This paper proposes the mixture model in Fig. 2 to incorporate a broader class of computational units that may be subject to hardware and calibration errors and characterizes the performance of ANT by the fidelity of these units.

3. PROBLEM DEFINITION

Let data D be generated from random variable X, $(X \to D)$ and let the computational purpose of a system be $D \to X$ with the *loss function* $\ell : \mathbb{R}^2 \to \mathbb{R}^+$ measuring the performance. The mixture model discussed in Section 3.1 is a generalization of the additive noise model of [3, 6, 7], Fig. 1 (a), that incorporates fidelity-ordered main, calibration and failure statistics.

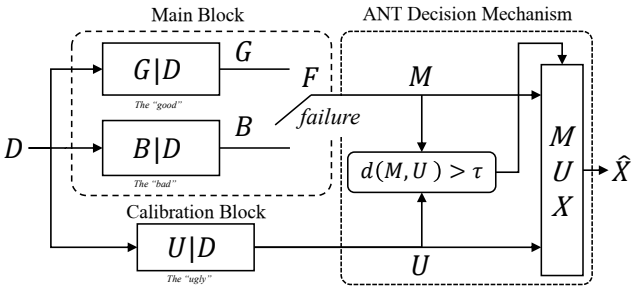


Fig. 2: Generalized model for ANT architecture

3.1. A Generalized Model for ANT

The ANT architecture is modeled as consisting of *three* computational units, henceforth called blocks: a main block that comprises a *good* block producing an outcome G and a bad one that produces B, which models intermittent failure of the main computational path, and a calibration, or so-called ugly, block producing U, modeling the lower-fidelity alternative, should it be decided that the main block has failed, as seen in Fig. 2. Intuitively, the main block produces $M = G$, when there is no hardware failure, and produces $M = B$ when there is a failure. The ANT decision mechanism can use the calibration block U to detect hardware failures and bypass them by switching between M and U.

Formally, $X \to D \to (M, U) \equiv (\{G, B\}, U)$ and the ANT decision rule, denoted by $\delta^{ANT}(\cdot)$, operates on the branch outcomes, (M, U), to make a decision between them: $\delta^{ANT}(M, U) \in \{M, U\} \equiv \{G, B, U\}$. The block with *lower expected loss* is understood to have *higher fidelity*:

$$\mathbb{E}\ell\,(G, X) < \mathbb{E}\ell\,(U, X) < \mathbb{E}\ell\,(B, X) \tag{1}$$

Here $\mathbb{E}\cdot$ denotes the expectation operator, defined over the triplet $(\Omega, \mathscr{F}, \mathbb{P})$, where we allow the random variables to be real-valued: $\Omega = \mathbb{R}$, $\mathscr{F} = \mathscr{B}(\mathbb{R})$ is the Borel σ-algebra and $\mathbb{P}(\cdot)$ is the probability measure.

The triplet (G, B, U) is conditioned on the hidden variable X and the conditional probability density functions $p_{G|X}$, $p_{B|X}$ and $p_{U|X}$ represent the statistical characteristics of the main block, the main block under hardware failure and the calibration block respectively. These distributions represent the computational properties of respective blocks under uncertainties due to process, temperature and voltage variations, which are commonly unknown or too costly to model [1].

We allow the process of each block to be statistically independent, that is, given the hidden variable X, outcomes

G, B, U are conditionally independent from one another: $\forall C \in \{G, B, U\}$, $C \leftrightarrow X \leftrightarrow \{G, B, U\} \setminus C$, equivalently, $p_{GBU|X} = p_{G|X} p_{B|X} p_{U|X}$, [11]. This assumption is similar to the independence of estimation error and hardware error in [3,6] and to the independence of noise in different branches in [7].

The generalized model for ANT allows the main block to *switch modes of operation* back and forth between the good block and the bad block. Letting an independent Bernoulli random variable of parameter $p \in (0, 1)$, denoted by $F \sim \mathcal{B}(p)$, determine the hardware failure, the following mixture model characterizes the main block:

$$M = G\bar{F} + BF$$

Here, $\bar{F} = 1 - F$ and a failure happens when $F = 1$.

The purpose of ANT is to utilize the calibration random variable U to determine whether a failure $(F = 1)$ happens, or not $(F = 0)$, and bypass the main block with the calibration block when it does. If $p_{G|X}$ and $p_{B|X}$, were known a priori, likelihood ratio would provide the sufficient statistics for testing whether a failure has occurred or not, and the Neyman-Pearson rule could be built upon it [10]. Instead, the ANT architecture *builds* statistics using the pair $(M, U) \equiv (\{G, B\}, U)$. In Section 3.2, we introduce the ANT decision rule on an arbitrary metric space and define a measure of performance as the regret with respect to the optimal-yet-unattainable oracle decision rule.

3.2. Performance Criterion for ANT

ANT builds decision statistics from the pair (M, U) to test whether $M \sim G$ or $M \sim B$ and bypass the main block when $M = B$. Let $d(\cdot, \cdot)$ be a distance measure defined on $\mathbb{R}$, satisfying the axioms in [12], a general ANT decision rule has the following form:

$$\delta^{ANT}(M, U) = \begin{cases} M & \text{if} \quad d(M, U) \leq \tau \\ U & \text{if} \quad d(M, U) > \tau \end{cases}$$

Intuitively, ANT decision rule "favors" the main computational unit when it passes a "calibration check", otherwise it uses the calibration unit to bypass the main unit that is "flagged" with hardware failure.

Now consider an *oracle* that has access to reliable information on when a hardware failure, F, occurs. Such an oracle minimizes its loss via the following decision rule:

$$\delta^{O}(M, U|F = f) = \begin{cases} M & \text{if} \quad f = 0 \\ U & \text{if} \quad f = 1 \end{cases}$$

In practice, information on F is not easily, if at all, available. However, it serves as a useful benchmark for measuring the performance of ANT. This paper proposes a conservative measure of performance by defining the *regret* of ANT as the expected loss suffered from using the ANT decision rule, δ^{ANT}, against that of the oracle decision rule δ^{O}:

$$R^{ANT}(\tau) = \mathbb{E}\ell\left(\delta^{ANT}(M, U), X\right) - \mathbb{E}\ell\left(\delta^{O}(M, U|F), X\right)$$

A more intuitive form for $R^{ANT}(\tau)$, follows from the independence of F and the total law of probability, [11]:

Proposition 1. *For any triplet (G, B, U) of computational units, the regret of ANT satisfies:*

$$R^{ANT}(\tau) = \bar{p} R_{UG} \Phi_d^{GU}(\tau) + p R_{BU} F_d^{BU}(\tau). \quad (2)$$

Here, $\bar{p} = 1 - p$, $R_{\alpha\beta} \triangleq \mathbb{E}\ell(\alpha, X) - \mathbb{E}\ell(\beta, X)$, where $(\alpha, \beta) \subset \{G, B, U\}$, $\Phi_d^{GU} \triangleq \mathbb{P}(d(G, U) > \tau)$ and $F_d^{BU} \triangleq \mathbb{P}(d(B, U) \leq \tau)$, when the distance metric is known from context, we drop the subscript. The regret in (2) shows that when ANT "misses" a hardware failure, which happens with probability $F_d^{BU}(\tau)$, it is incurred a regret of pR^{BU}. Similarly, with probability $\Phi_d^{GU}(\tau)$, ANT raises a "false alarm" and switches back to U at regret $\bar{p}R^{GU}$. Let's remark that R^{UG} and R^{BU} are functionals of $\ell(\cdot, \cdot)$, where $\Phi_d^{GU}(\tau)$ and $F_d^{BU}(\tau)$ are functionals of $d(\cdot, \cdot)$. In Section 4, we explore the connection between these functionals and quantify a fidelity based characterization of the regret.

4. FIDELITY-BASED CHARACTERIZATION

The regret of ANT is a mixed functional of the distance measure $d(\cdot, \cdot)$ used to build the decision statistics and the loss function $\ell(\cdot, \cdot)$ that determines the fidelity of a computational unit. On a Hilbert space, $\mathcal{H}$, the distance measure is given by:

$$d(M, C) = \|M - C\|$$

where $\|\cdot\|$ is the norm associated with $\mathcal{H}$, [12]. This section explores the fundamental limits for the regret of ANT for fidelities defined by any $\mathcal{C}$-bi-Lipschitz loss function on a Hilbert space, $\mathcal{H}$. That is, $\forall \{M, C\} \subset \{G, B, U\}$:

$$\frac{1}{\mathcal{C}} \|M - C\| \leq |\ell(M, X) - \ell(C, X)| \leq \mathcal{C} \|M - C\| \quad (3)$$

In Section 4.1, we discuss the necessary conditions for ANT to achieve the performance of an oracle decision rule.

4.1. Necessary Conditions for Optimal ANT

If $\exists \tau : R^{ANT}(\tau) = 0$, then ANT is optimal, that is, it operates with no-regret. The fidelity ordering in (1) yields that:

$$R^{ANT}(\tau) = 0 \iff \Phi_d^{GU}(\tau) = F_d^{BU}(\tau) = 0. \quad (4)$$

This follows from positivity of p, $\bar{p}$, R^{GU} and R^{BU} and it yields the following statistical necessary condition and its fidelity-based sufficient counterpart:

Proposition 2. *A necessary condition for* (4) *is* $\mathbb{E}\|U-G\| < \mathbb{E}\|B-U\|$*, which is satisfied* $\forall(G,B,U)$ *such that* $\mathbb{E}\ell(G,X) < \mathbb{E}\ell(U,X) < \frac{1}{2\mathcal{C}^2+1}\mathbb{E}\ell(B,X)$.

The proof is outlined in Section 5 and follows from (3) and the triangle inequality, [12]. In Section 4.2, we propose fidelity conditions under which the regret of ANT is universally bounded.

4.2. Sufficient Conditions for ε-ANT

We call the setup where $\exists\,(\varepsilon,\tau)$ such that, $R^{ANT}(\tau)\leq\varepsilon$, an ε-ANT. First, we make a few observations that follow from (3) and triangle inequality:

1. $\mathbb{E}\|B-U\| \geq \frac{1}{\mathcal{C}}R^{BU}$ and $\mathbb{E}\|G-U\| \leq \mathcal{C}\Sigma^{GU}$, where $\Sigma^{GU}\triangleq\mathbb{E}\ell(G,X)+\mathbb{E}\ell(U,X)$.
2. Chernoff Bound, [11]: $\log\Phi^{GU}(\tau) \leq -\frac{\left(\tau-\mathcal{C}\Sigma^{GU}\right)^2}{3\mathcal{C}\Sigma^{GU}}$ and $\log F^{BU}(\tau)\leq-\frac{\left(\mathcal{C}\tau-R^{BU}\right)^2}{3\mathcal{C}R^{BU}}$.

Here, $\Sigma^{GU}=\mathbb{E}\ell(G,X)+\mathbb{E}\ell(U,X)$ and we further define $\Delta\triangleq R^{BU}-\Sigma^{GU}$. These observations allow us to construct the main result of this paper; the sufficient condition under which ε-ANT exists.

Theorem 1. *For any triplet* (G,B,U) *that satisfies* $\Sigma^{GU}>\frac{1}{\bar{p}}$ *and* $R^{BU}>1/p$*, if*

$$\exp\frac{\Delta^2}{3\Sigma^{GU}\left(\mathcal{C}+\frac{\bar{p}}{\mathcal{C}p}\right)}+\exp\frac{\Delta^2}{3R^{BU}\left(\frac{1}{\mathcal{C}}+\frac{\mathcal{C}p}{\bar{p}}\right)}>\varepsilon \quad (5)$$

then, $\exists\tau$ *such that* $R^{ANT}(\tau)\leq\varepsilon$.

The first term on the left hand side of (5), controls the regret due to "false alarm" and the second one controls that of "miss". The proof of Theorem 1 is outlined in Section 5. We remark that the Chernoff bound is not generally sharp, however, it lends the necessary tool for our fidelity-based analysis to be universal, which we demonstrate in Section 6.

5. PROOFS

Outline for the Proof of Proposition 2. The necessity follows from $\Phi^{GU}(\tau)=F^{BU}(\tau)=0\Rightarrow\mathbb{E}\|U-G\|<\tau<\mathbb{E}\|B-U\|$. The sufficient condition follows from the linearity of expectation applied to triangle inequality and (3) for $\|G-U\|$, and non-negativity of the loss function and equations (1), (3) for $\|B-U\|$. □

Outline of the Proof of Theorem 1. The main idea is to use the Chernoff bound, [11], to upper-bound the regret $R^{ANT}(\tau)$, which is in terms of $\mathbb{E}\|U-G\|$ and $\mathbb{E}\|B-U\|$, then use observation 1 and monotonicity of the Chernoff bound (with respect to the expectation) to postulate the form in observation 2. Finally, observing that the minimizer is a logarithmic Lambert function, an upper bound on the minimum yields the theorem. □

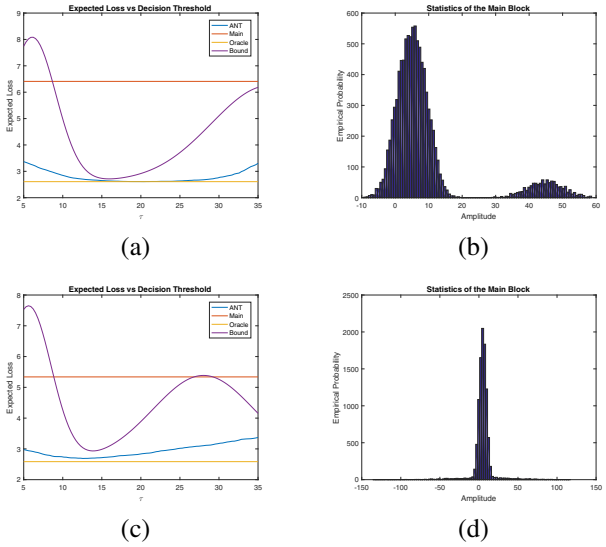


Fig. 3: Universal bounds vs performance of ANT for different error statistics: (a-b) for bias introducing hardware failure, (c-d) burying noise (high variance) hardware failure

6. EXPERIMENTS

We propose a Gaussian mixture with $G\sim\mathcal{N}(X,\sigma_G)$, $U\sim\mathcal{N}(X,\sigma_U)$ and the bad-block either introducing a bias $B\sim\mathcal{N}(X+\mu_B,\sigma_B)$ or introducing a very large variance noise. We compare the performance of ANT to that of the *oracle* over the range of τ values and demonstrate that the Chernoff bounds that we propose in observation 2, indicate accurately when the performance of ANT is optimal.

Experiment specifications are as follows: Figure 3 (a)-(b) demonstrate the setup, where the bias that the "bad" block introduces is the main source of distortion as, $X\sim Unif(0,10)$, $\sigma_G=10$, $\sigma_B=\sigma_U=20$ with $\mu_B=40$. Figure 3 (c)-(d) illustrates the case, where the "bad" block has very large variance: $\sigma_G=10$, $\sigma_U=15$, yet $\sigma_B=1.5e+3$. As expected from Proposition 2 as σ_B decreases, the performance of ANT deteriorates.

A key observation is that minimizing the upper bound on $R^{ANT}(\tau)$ yields a τ value that is close to the true and statistics-dependent optimal threshold value. This idea further captures, albeit with less accuracy, the region of τ values for which near-optimal ANT performance is maintained.

7. CONCLUSION

In this paper, we proposed a generalized model for the ANT architecture and derived fidelity-based conditions that are universal over the statistical properties of the underlying computational units. We showed that under these conditions, the performance of ANT is guaranteed to be quantifiably close to that of an oracle decision rule that can not be attained.

8. REFERENCES

[1] Naresh R Shanbhag, Rami A Abdallah, Rakesh Kumar, and Douglas L Jones, "Stochastic computation," in *Design Automation Conference (DAC), 2010 47th ACM/IEEE*. IEEE, 2010, pp. 859–864.

[2] Rajamohana Hegde and Naresh R Shanbhag, "Soft digital signal processing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 9, no. 6, pp. 813–823, 2001.

[3] Sai Zhang and Naresh R Shanbhag, "Embedded algorithmic noise-tolerance for signal processing and machine learning systems via data path decomposition," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3338–3350, 2016.

[4] Rami A Abdallah and Naresh R Shanbhag, "An energy-efficient ecg processor in 45-nm cmos using statistical error compensation," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 11, pp. 2882–2893, 2013.

[5] Eric P Kim, Daniel J Baker, Sriram Narayanan, Douglas L Jones, and Naresh R Shanbhag, "Low power and error resilient pn code acquisition filter via statistical error compensation," in *Custom Integrated Circuits Conference (CICC), 2011 IEEE*. IEEE, 2011, pp. 1–4.

[6] Eric P Kim and Naresh R Shanbhag, "Statistical analysis of algorithmic noise tolerance," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2731–2735.

[7] Daewon Seo and Lav R Varshney, "Information-theoretic limits of algorithmic noise tolerance," in *Rebooting Computing (ICRC), IEEE International Conference on*. IEEE, 2016, pp. 1–4.

[8] Mehmet A Donmez, Maxim Raginsky, Andrew C Singer, and Lav R Varshney, "Cost-performance tradeoffs in unreliable computation architectures," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 215–219.

[9] Jun Won Choi, Byonghyo Shim, Andrew C Singer, and Nam Ik Cho, "Low-power filtering via minimum power soft error cancellation," *IEEE Transactions on Signal Processing*, vol. 55, no. 10, pp. 5084–5096, 2007.

[10] H Vincent Poor, *An introduction to signal detection and estimation*, Springer Science & Business Media, 2013.

[11] Bruce Hajek, *Random processes for engineers*, Cambridge university press, 2015.

[12] W. Rudin, *Functional Analysis*, International series in pure and applied mathematics. McGraw-Hill, 1991.