A fully-analogue, universal primitive for linear algebra operations using resistive memories

Piergiulio Mannocci, *Graduate Student Member, IEEE,* Giacomo Pedretti, *Member, IEEE,* Elisabetta Giannone, Enrico Melacarne, Zhong Sun, *Member, IEEE,* and Daniele Ielmini, *Fellow, IEEE*

***Abstract*—The increasing demand for data-intensive computing applications, such as artificial intelligence (AI) and more specifically machine learning (ML), raises the need for novel computing hardware architectures capable of massive parallelism in performing core algebraic operations. Among the new paradigms, in-memory computing (IMC) with analogue devices is attracting significant interest for its large-scale integration potential, together with unrivaled speed and energy performance. Here, we present a fully-analogue, universal primitive capable of executing linear algebra operations such as regression, generalized least-square minimization and linear system solution with and without preconditioning. We study the impact of the main circuit parameters on accuracy and bandwidth with analytical closed-form expressions and SPICE simulations. Finally, the scaling challenges due to parasitic resistance/capacitance and the possible solutions to overcome these effects are discussed.**

***Index Terms*—In-memory computing, resistive memory, hardware accelerator, linear regression, linear systems**

I. Introduction

An ever-increasing number of computing tasks make use of linear regression and matrix inversion in their core primitives. Both these operations are costly in terms of energy and time, partly because conventional computing systems are still based on von Neumann architecture, where data must be continuously shuttled between the memory and the processing unit [1]. While optimized digital hardware, such as the tensor processing unit (TPU) [2], has been developed to handle massive matrix computation, novel computing paradigms are needed to radically overcome the limits of the von Neumann architecture. Among these new concepts, in-memory computing (IMC) has gained momentum by directly performing computation in situ within the memory [3]. IMC is particularly suited for combination with emerging memory devices or *memristors*, such as the resistive-switching random access memory (RRAM), which can be integrated in high-density crossbar arrays [4]. IMC can theoretically achieve unrivaled speed, energy [3] and information capacitance [5] when compared with digital systems. Several demonstrations of IMC have been reported, including matrix-vector multiplication (MVM) in neural networks [6], [7], image processing [8], optimization problems [9], [10]. In particular, the iterative solution of linear systems [11], [12] has attracted interest for the improved speed and energy efficiency, however the need for a digital overhead may compensate the advantages IMC parallelism.

On the other hand, analogue computing as a concept dates back to the Greek civilization [13], and has been studied by many throughout history including Kirchhoff and Thomson [14]. After an interest decline in the first half of the 20th century due to the rise of digital computing, the works of Carver Mead [15] revitalized the field by identifying analogue computing as a suitable candidate for post-Moore's Law computing. In recent years, fully-analogue circuits for one-step computation of linear systems [16], [17] have been both theoretically and experimentally demonstrated, attracting interest for their inherent speed and reduced computational complexity. However, a detailed analysis of stability and bandwidth is necessary to fully assess their feasibility and currently lacking.

Here, we present a detailed study of a novel circuit for linear regression [18] to assess its accuracy and bandwidth. Various versions of the same circuit are shown for generalized least-square minimization and linear-system solution. We show that the linear-system circuit has an improved stability with respect to previous schemes [16], as well as being amenable to preconditioning to arbitrarily control the bandwidth. These circuits may be regarded as a universal primitive for IMC-based linear algebra operations. Finally, scaling challenges are discussed in terms of inherent parasitic resistance and capacitance.

In the following, we adopt the Householder notation [19], where bold capital letters $\mathbf{A}, \mathbf{B}$ denote matrices. $\mathbf{A}^T$ is the transpose of $\mathbf{A}$, $\|\cdot\|_p$ is the vector p-norm and $|||\cdot|||_p$ the induced operator p-norm. A Hermitian positive (negative) semidefinite matrix satisfies $\mathbf{A} \succeq 0$ ($\mathbf{A} \preceq 0$) and its singular values are $\sigma_1(\mathbf{A}) \geq \ldots \geq \sigma_n(\mathbf{A})$.

II. Circuit structure and steady-state operation

Figure 1a shows the proposed circuit schematic for the least-square linear regression solver, where the regressors $\mathbf{X} \in \mathbb{R}^{n \times m}$ are mapped in crosspoint arrays using a reference conductance G_0, *i.e.* $G_{X,ij} = G_0 X_{ij}$. The dependent variable y is mapped onto the input current array i_y in a similar fashion, $i_{y,i} = i_0 y_i$ where i_0 is a reference current. Finally, the output of amplifiers A_1 is locally fed back to the input by identical $G_f = G_0 f$ conductances. The state equations of the circuit can be written as:

$$\begin{cases} v_\varepsilon = -G_f^{-1}(G_0 \mathbf{X} v_\beta + i_y) & (1) \\ G_0 \mathbf{X}^T v_\varepsilon = 0 & (2) \end{cases}$$

P. Mannocci, E. Giannone, E. Melacarne and D. Ielmini are with Politecnico di Milano, Milano, Italy.

G. Pedretti is with Hewlett Packard Labs, Milpitas, CA, USA.

Z. Sun is with Peking University (PKU), Beijing, China.

Correspondence should be addressed to piergiulio.mannocci@polimi.it and daniele.ielmini@polimi.it.

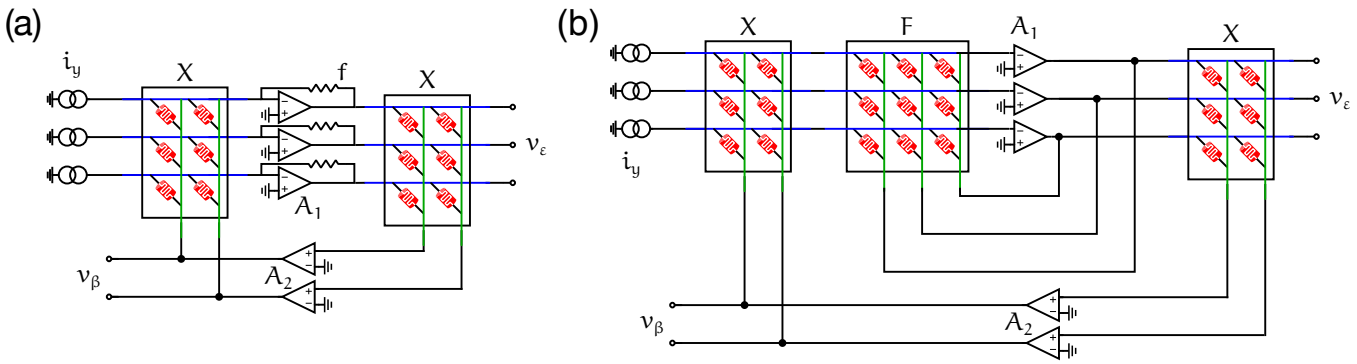

Fig. 1. Circuit schematic for performing (a) least squares (linear) regression and (b) generalized least squares regression. When an input is applied by the i_y generators, the output v_β provides the regression coefficients, whereas the residuals are provided by v_ε. Both circuits can also be employed as linear system solvers when $\mathbf{X}$ is square, with (b) allowing to map a preconditioning matrix in the $\mathbf{F}$ array.

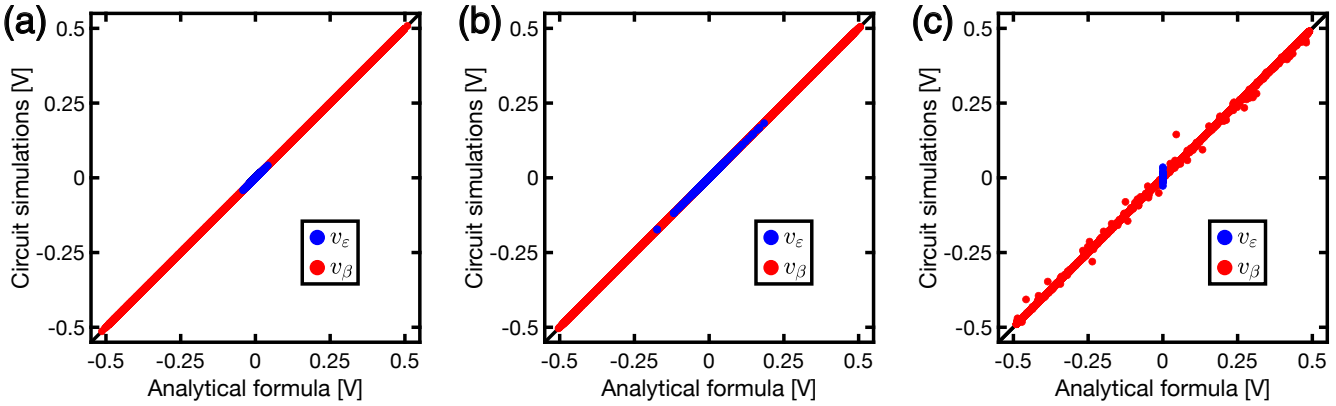

Fig. 2. Comparison of computed solution and analytical formula for 1000 SPICE simulations of (a) linear regression circuit on a 20×10 data matrix, (b) generalized least squares on a 20×10 data matrix and (c) linear system solution on a 10×10 matrix. Red dots represent the output voltages v_β of the A_2 set, which map the solution β, whereas blue dots represent the error output voltages v_ε of the A_1 set, which map the residual ε.

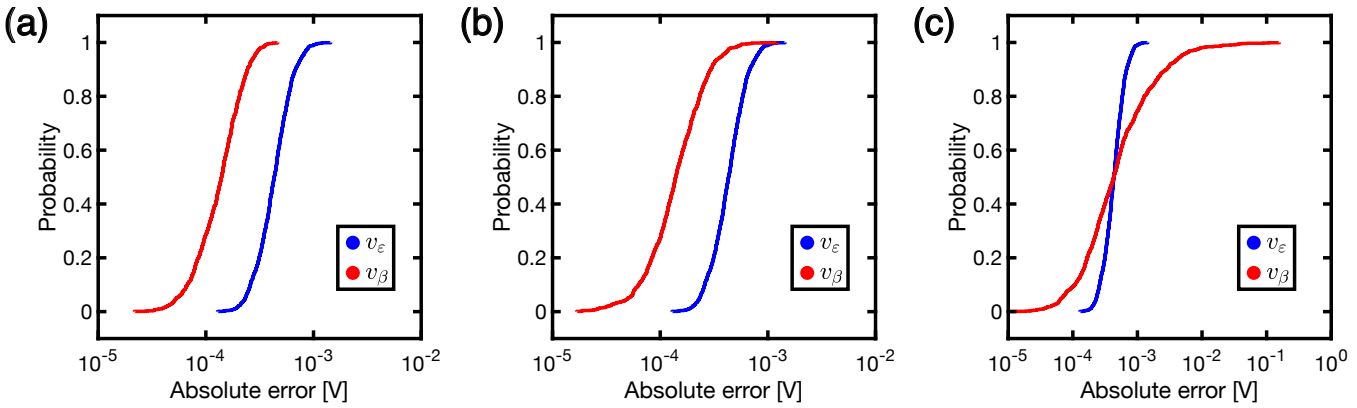

Fig. 3. Cumulative distribution functions of the absolute error for simulations of Fig. 2. (a) CDF of the absolute error on v_β, namely $\|v_\beta - v_{\beta,id}\|_2$ (red) and on v_ε, namely $\|v_\varepsilon - v_{\varepsilon,id}\|_2$, for the linear regression simulations in Fig. 2a. (b) CDF of the absolute error on v_β (red) and v_ε (blue) for the generalized linear regression simulations in Fig. 2b. (c) CDF of the absolute error on v_β (red) and v_ε (blue) for the linear system solution simulations in Fig. 2c. While the absolute error on v_ε is mostly independent of the given circuit configuration, the error on v_β shows a wider spread in the linear system solver configuration owing to the worse conditioning of the chosen linear system matrices.

where Eqs. 1 and 2 are the current balances at the input of A_1 and A_2 respectively. By substituting Eq. 1 in Eq. 2, we get the stationary transfer for ideal feedback conditions given by:

$$\begin{aligned} v_{\beta,id} &= -G_0^{-1}(\mathbf{X}^T f^{-1}\mathbf{X})^{-1}\mathbf{X}^T f^{-1} i_y \\ &= -G_0^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T i_y = -G_0^{-1}\mathbf{X}^+ i_y \end{aligned} \tag{3}$$

where $\mathbf{X}^+$ is the Moore-Penrose pseudoinverse of $\mathbf{X}$. By substituting Eq. 3 into Eq. 1, we obtain:

$$v_{\varepsilon,id} = -G_0^{-1} f^{-1}(\mathbf{I} - \mathbf{P}) i_y = -G_0^{-1} f^{-1}\mathbf{M} i_y \tag{4}$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the projection matrix and $\mathbf{M} = \mathbf{I} - \mathbf{P}$ is the residual maker matrix. The output vector $v_{\beta,id}$ can be viewed as the least-square regression of the equation

$\mathbf{X}\beta = y$, while $v_{\varepsilon,id}$ is the residual vector ε which satisfies $\mathbf{X}\beta + \varepsilon = y$. Fig. 2a shows the correlation plot of v_β and v_ε obtained from a SPICE circuit simulation as a function of the results of Eqs. 3 and 4 for a random regression problem of size 20×10, demonstrating the accuracy of our analytical model. Correspondingly, Fig. 3a shows the cumulative distribution function (CDF) of the absolute error for v_β, $\|v_\beta - v_{\beta,id}\|_2$, and v_ε, $\|v_\varepsilon - v_{\varepsilon,id}$, where the operational amplifiers' gain was set to 100 dB. The circuit of Fig. 1a can be generalized as shown in Fig. 1b, where the scalar conductance G_f is replaced by the conductance array $\mathbf{G_F} = G_0\mathbf{F}$, by following the scheme of the IMC-based linear-system solver [16], [17]. The output voltage now reads:

$$v_{\beta,id} = -G_0^{-1}(\mathbf{X}^T\mathbf{F}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{F}^{-1}i_y \tag{5}$$

while the residual is given by:

$$\begin{aligned} v_{\varepsilon,id} &= -G_0^{-1}\mathbf{F}^{-1}(\mathbf{X} - \mathbf{X}(\mathbf{X}^T\mathbf{F}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{F}^{-1})i_y \\ &= -G_0^{-1}\mathbf{F}^{-1}(\mathbf{I} - \tilde{\mathbf{P}})i_y = -G_0^{-1}\mathbf{F}^{-1}\tilde{\mathbf{M}}i_y \end{aligned} \tag{6}$$

which solves the generalized least-square regression of the input current i_y and the regressors $\mathbf{G_X}$, where matrices $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{M}} = \mathbf{I} - \tilde{\mathbf{P}}$ are the projection matrix and the residual maker matrix for the generalized least-square problem. The matrix $\mathbf{G_F}$ can be used to map the conditional variance of the error, *i.e.* the covariance matrix $Cov[\varepsilon|\mathbf{X}] = \mathbf{F}$. From a statistical standpoint, the introduction of the covariance matrix in the estimator adds information about correlation among the dataset samplings, rather than treating them as uncorrelated, resulting in an improved regression estimate. Fig. 2b shows the correlation plot of v_β and v_ε from a SPICE circuit simulation as a function of the results of Eqs. 5 and 6 for a random generalized least-square regression problem of size 20×10, demonstrating the accuracy of our analytical model. Correspondingly, Fig. 3b shows the cumulative distribution function (CDF) of the absolute error for v_β, $\|v_\beta - v_{\beta,id}\|_2$, and v_ε, $\|v_\varepsilon - v_{\varepsilon,id}$, where the operational amplifiers' gain was set to 100 dB.

For square matrices, *i.e.* $\mathbf{X} \in \mathbb{R}^{n\times n}$, both circuits in Fig. 2 solve the linear system $\mathbf{X}\beta = y$, resulting in the state equation:

$$v_{\beta,id} = -G_0^{-1}\mathbf{X}^{-1}i_y$$

with $v_\varepsilon = 0$. In addition, the circuit in Fig. 2b allows matrix preconditioning by properly mapping the $\mathbf{F}$ array. Fig. 2c shows the correlation plot of v_β and v_ϵ from a SPICE circuit simulation as a function of the analytical solution for a random 10×10 matrix, demonstrating their accuracy. Correspondingly, Fig. 3c shows the cumulative distribution function (CDF) of the absolute error for v_β, $\|v_\beta - v_{\beta,id}\|_2$, and v_ε, $\|v_\varepsilon - v_{\varepsilon,id}$, where the operational amplifiers' gain was set to 100 dB.

Thanks to the negative feedback, all the presented operations are efficiently executed in one computational step.

III. Transient Analysis

THE circuit of Fig. 1 was modeled by the block diagram of Fig. 4. Here, the amplifier sets A_1 and A_2 are described by scalar matrices $\mathbf{A_1} \in \mathbb{R}^{n\times n}$ and $\mathbf{A_2} \in \mathbb{R}^{m\times m}$, where each diagonal element contains the transfer function of the corresponding amplifier in the Laplace domain, namely $\alpha_1(s)$ and $\alpha_2(s)$ respectively. Summation nodes are described by diagonal matrices $\mathbf{U_1} \in \mathbb{R}^{n\times n}$, $\mathbf{U_2} \in \mathbb{R}^{m\times m}$ whereas $\mathbf{X} \in \mathbb{R}^{n\times m}$, $\mathbf{X}^T \in \mathbb{R}^{m\times n}$, $\mathbf{F} \in \mathbb{R}^{n\times n}$ describe the transfer of A_2 output to A_1 input, A_1 output to A_2 input and A_1 to itself, respectively. Referring to the circuit of Fig. 1b, the summation matrices can be written as [16]:

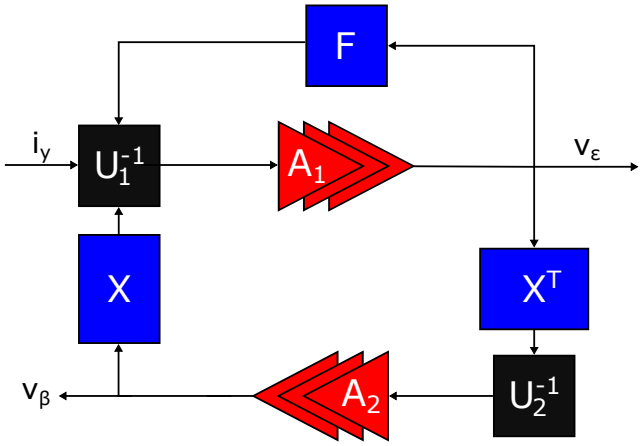

Fig. 4. Block diagram of the circuit of Fig. 1b. Blocks $\mathbf{F}$, $\mathbf{X}$ and $\mathbf{X}^T$ are conductance arrays. Blocks A_1 and A_2 are sets of operational amplifiers and blocks $\mathbf{U_1}^{-1}$ and $\mathbf{U_2}^{-1}$ are summing nodes. The input current i_y is applied to the summing node of A_1, while the output voltages of A_1 and A_2 correspond to v_ε and v_β respectively.

$$U_{1,ii} = \sum_{j=1}^{n} X_{ij} + \sum_{j=1}^{n} F_{ij}$$

$$U_{2,ii} = \sum_{j=1}^{m} X_{ij}^T = \sum_{j=1}^{m} X_{ji}.$$

From Fig. 4, v_β can be written as a function of v_ϵ and vice versa, namely:

$$\begin{cases} v_\beta = \mathbf{A_2}\mathbf{U_2}^{-1}\mathbf{X}^T v_\varepsilon \\ v_\varepsilon = \mathbf{A_1}\mathbf{U_1}^{-1}(G_0^{-1}i_y + \mathbf{X}v_\beta + \mathbf{F}v_\varepsilon) \end{cases}$$

yielding the following expressions for v_β and v_ϵ:

$$v_\varepsilon = (\mathbf{U_1}\mathbf{A_1}^{-1} - \mathbf{X}\mathbf{A_2}\mathbf{U_2}^{-1}\mathbf{X}^T + \mathbf{F})^{-1}G_0^{-1}i_y \tag{7}$$

$$\begin{aligned} v_\beta = -\ &(-\mathbf{A_1}^{-1}\mathbf{A_2}^{-1}\mathbf{X}^T\mathbf{U_1}(\mathbf{X}^T)^+\mathbf{U_2} \\ &+ \mathbf{A_2}^{-1}\mathbf{X}^T\mathbf{F}(\mathbf{X}^T)^+\mathbf{U_2}\ + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T G_0^{-1}i_y \end{aligned} \tag{8}$$

which provide the input-output transfer functions of the circuit.

A. Static error

Given the single-pole transfer functions of the operational amplifiers $\alpha_1(s) = -\frac{\alpha_1}{1+s\tau_1}$ and $\alpha_2(s) = +\frac{\alpha_2}{1+s\tau_2}$ with $\alpha_1, \alpha_2 > 0$, we can define matrix $\mathbf{E}$ as:

$$\mathbf{E} = \frac{1}{\alpha_1\alpha_2}\mathbf{X}^T\mathbf{U_1}(\mathbf{X}^T)^+\mathbf{U_2} + \frac{1}{\alpha_2}\mathbf{X}^T\mathbf{F}(\mathbf{X}^T)^+\mathbf{U_2}$$

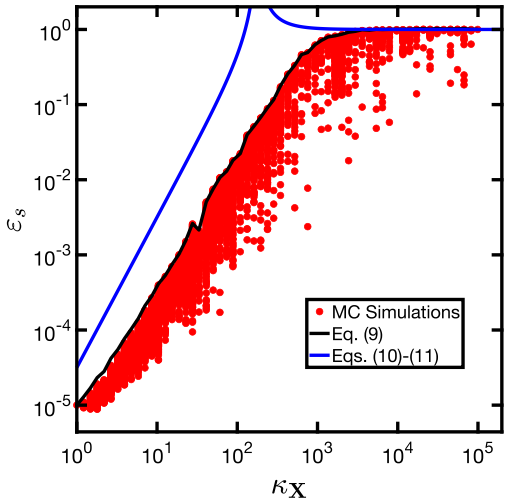

Fig. 5. Relative error of the calculated v_β as a function of the condition number $\kappa_{\mathbf{X}}$ for a 10×10 linear system with no preconditioning ($\mathbf{F} = \mathbf{I}$). Each point is the result for a linear system with random input and matrix. The theoretical limits of Eqs. (9) and (10)-(11) are also shown.

so that Eq. (8) can be rewritten at steady state ($s = 0$) as:

$$v_\beta = -(\mathbf{E} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T G_0^{-1} i_y = -(\mathbf{I} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{E})^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T G_0^{-1} i_y$$

From Eq. (3), we can evaluate the ideal output for $\alpha_1 = \alpha_2 = \infty$, hence $\mathbf{E} = 0$, which yields $v_{\beta,id} = -G_0^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T i_y$. The relative static error ε_s due to the finite amplifier gain thus obeys:

$$\varepsilon_s = \frac{\|v_\beta - v_{\beta,id}\|_2}{\|v_{\beta,id}\|_2} \leq \left\|\left|(\mathbf{I} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{E})^{-1} - \mathbf{I}\right|\right\|_2 \quad (9)$$

$$= \left\|\left|(\mathbf{I} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{E})^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{E}\right|\right\|_2 \leq \frac{\left\|\left|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{E}\right|\right\|_2}{1 - \left\|\left|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{E}\right|\right\|_2} \quad (10)$$

where:

$$0 \leq \left\|\left|(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{E}\right|\right\|_2 \leq \kappa_{\mathbf{X}}^2 \frac{n}{\alpha_1\alpha_2}\left(1 + \frac{\alpha_1 + \sqrt{n}}{\sqrt{n}}\left\|\left|\mathbf{F}\right|\right\|_2\right) \quad (11)$$

with $\kappa_{\mathbf{X}}$ the ℓ_2-norm condition number of $\mathbf{X}$, and we used $\left\|\left|\mathbf{U_2}\right|\right\|_2 = \left\|\left|\mathbf{X}\right|\right\|_1$, $\left\|\left|\mathbf{U_1}\right|\right\|_2 \leq \left\|\left|\mathbf{X}\right|\right\|_\infty + \left\|\left|\mathbf{F}\right|\right\|_\infty$. Fig. 5 shows the calculated ε_s for random 10×10 linear systems with variable condition number $\kappa_{\mathbf{X}}$. The maximum limit given by Eqs. (10)-(11) is extremely conservative when tested against uniformly-random dense matrices, and shows a quadratic dependence on the condition number $\kappa_{\mathbf{X}}$ and a linear dependence on $\sigma_1(\mathbf{F})$. For sufficiently large α_1, the impact of this gain on the relative error is negligible, thus allowing to relax the gain-bandwidth trade-off on the A_1 set during the design phase.

B. Stability

To assess the stability of the circuits in Fig. 1, the system poles must be computed from the transfer functions of Eqs. (7) and (8). Eq. (8) can be rewritten as:

$$\begin{aligned} v_\beta &= -\mathbf{U_2}^{-1}\mathbf{X}^T(-\mathbf{A_1}^{-1}\mathbf{A_2}^{-1}\mathbf{X}^T\mathbf{U_1} + \mathbf{A_2}^{-1}\mathbf{X}^T\mathbf{F} \\ &\quad + \mathbf{X}^T\mathbf{X}\mathbf{U_2}^{-1}\mathbf{X}^T)^{-1}\mathbf{X}^T G_0^{-1} i_y \\ &= -\mathbf{U_2}^{-1}\mathbf{X}^T(-\mathbf{A_1}^{-1}\mathbf{A_2}^{-1}\mathbf{U_1} + \mathbf{A_2}^{-1}\mathbf{F} \\ &\quad + \mathbf{X}\mathbf{U_2}^{-1}\mathbf{X}^T)^{-1} G_0^{-1} i_y \\ &= -\mathbf{U_2}^{-1}\mathbf{X}^T(-\mathbf{A_1}^{-1}\mathbf{A_2}^{-1} + \mathbf{A_2}^{-1}\mathbf{U_1}^{-1}\mathbf{F} \\ &\quad + \mathbf{U_1}^{-1}\mathbf{X}\mathbf{U_2}^{-1}\mathbf{X}^T)^{-1}\mathbf{U_1}^{-1} G_0^{-1} i_y \end{aligned}$$

which allows to obtain the characteristic equation for the circuit poles, namely [17], [18]:

$$(-\mathbf{A_1}^{-1}\mathbf{A_2}^{-1} + \mathbf{A_2}^{-1}\mathbf{U_1}^{-1}\mathbf{F} + \mathbf{U_1}^{-1}\mathbf{X}\mathbf{U_2}^{-1}\mathbf{X}^T)w = 0$$

For the sake of simplicity, the amplifiers may be considered identical apart from their respective signs, such that $\mathbf{A_1} = -\alpha(s)\mathbf{I}_n$, $\mathbf{A_2} = \alpha(s)\mathbf{I}_m$, $\alpha(s) \neq 0$, and thus:

$$\left(\frac{1}{\alpha(s)^2}\mathbf{I}_n + \frac{1}{\alpha(s)}\mathbf{U_1}^{-1}\mathbf{F} + \mathbf{U_1}^{-1}\mathbf{X}\mathbf{U_2}^{-1}\mathbf{X}^T\right)w = 0$$

This can be rewritten as an Hermitian Quadratic Eigenvalue Problem (QEP) [20], namely:

$$(\lambda^2\mathbf{I}_n + \lambda\mathbf{U_1}^{-\frac{1}{2}}\mathbf{F}\mathbf{U_1}^{-\frac{1}{2}} + \mathbf{U_1}^{-\frac{1}{2}}\mathbf{X}\mathbf{U_2}^{-1}\mathbf{X}^T\mathbf{U_1}^{-\frac{1}{2}})w = 0 \quad (12)$$

where we assumed $1/\lambda = \alpha(s)$. Considering a model single-pole amplifier, i.e $\alpha(s) = \frac{\alpha_0}{1+s\tau_0}$ with $\alpha_0 > 0$, then the poles of the circuit are given by:

$$p_i = \frac{1}{\tau_0}(\alpha_0\lambda_i - 1) \quad (13)$$

where λ_i are the nonzero solutions of the QEP. Notice that for $\mathbf{X} \in \mathbb{R}^{n\times m}$, Eq. (12) returns $2n$ solutions of which $n+m$ nonzero values represent the effective poles of the system. In the case of a linear system solver, where the two matrices are square, the QEP correctly returns $2n$ solutions.

For stability, all p_i should have negative real parts, namely $\Re(p) \leq 0$, which is always true if $\Re(\lambda) \leq 0$. Sufficient conditions are [20] (1) $\mathbf{I}_n \succ 0$, which is always satisfied, (2) $\mathbf{U_1}^{-\frac{1}{2}}\mathbf{F}\mathbf{U_1}^{-\frac{1}{2}} \succeq 0$ and (3) $\mathbf{U_1}^{-\frac{1}{2}}\mathbf{X}\mathbf{U_2}^{-1}\mathbf{X}^T\mathbf{U_1}^{-\frac{1}{2}} \succeq 0$, which is always true since $\mathbf{U_1}^{-\frac{1}{2}}\mathbf{X}\mathbf{U_2}^{-1}\mathbf{X}^T\mathbf{U_1}^{-\frac{1}{2}} = (\mathbf{U_1}^{-\frac{1}{2}}\mathbf{X}\mathbf{U_2}^{-\frac{1}{2}})(\mathbf{U_1}^{-\frac{1}{2}}\mathbf{X}\mathbf{U_2}^{-\frac{1}{2}})^T$ is a Gram matrix, hence always positive semi-definite. Condition (2) requires $\mathbf{F} \succeq 0$, which can be regarded as the only condition for stability. In the case of generalized least squares $\mathbf{F}$ is a covariance matrix which is always positive semidefinite. When performing linear system solution instead, the preconditioner $\mathbf{F}$ can be chosen to be positive semi-definite. As a result, the circuit in Fig. 1b is always stable for all practical cases, irrespective of the regressor matrix $\mathbf{X}$. Under these conditions, the roots λ of Eq. (12) are given by:

$$\lambda = \frac{-q_2(x) \pm \sqrt{q_2(x)^2 - 4q_1(x)q_3(x)}}{2q_1(x)} \quad (14)$$

where $q_1(x) = x^T\mathbf{I}x = 1$, $q_2(x) = x^T\mathbf{U_1}^{-\frac{1}{2}}\mathbf{F}\mathbf{U_1}^{-\frac{1}{2}}x$ and $q_3(x) = x^T\mathbf{U_1}^{-\frac{1}{2}}\mathbf{X}\mathbf{U_2}^{-1}\mathbf{X}^T\mathbf{U_1}^{-\frac{1}{2}}x$, with $\|x\|_2 = 1$ [20].

In the scalar case, $\mathbf{F} = f\mathbf{I}$. In the limit $f \to 0$, $q_2(x) \to f\tilde{q}_2(x)$ where $\tilde{q}_2(x) \neq 0$, thus $\lambda \simeq -\frac{1}{2}f\tilde{q}_2(x) \pm j\sqrt{q_3(x)}$. For small values of f, each pole pair is a complex conjugate

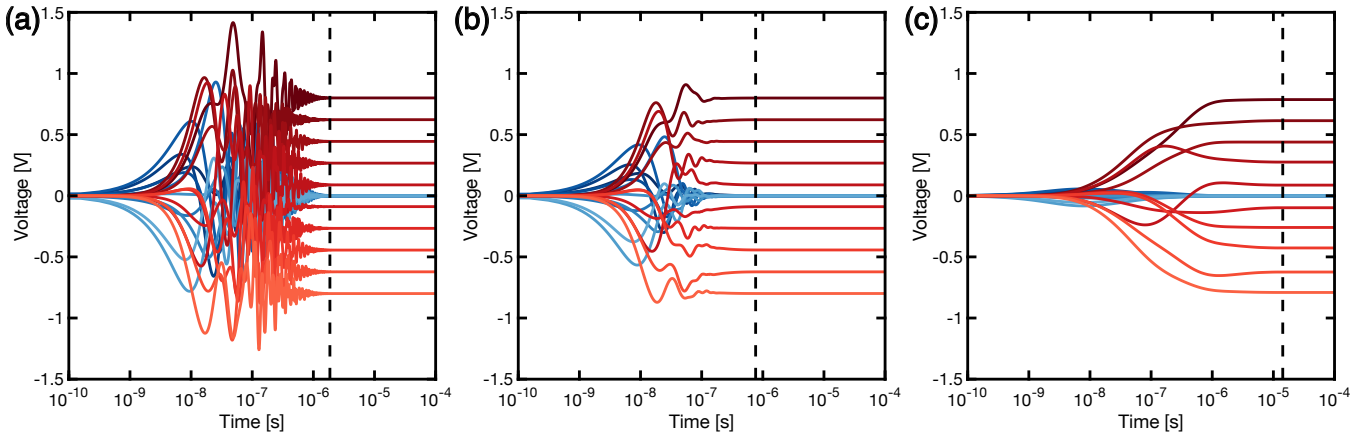

Fig. 6. Calculated time evolution of v_ε and v_β for a 10×10 linear system with $\kappa_{\mathbf{X}} = 27$ and different values of f for controlling the damping, namely (a) underdamped regime (f = 0.025), (b) optimally damped (f = 0.25) and (c) overdamped (f = 15.8). The dashed line highlights the solution time, defined as the inverse of the real part of the dominant pole, $t_{sol} = 1/\Re(p_{min})$. Damping causes an increase of the solution time and a reduction of the ringing effect.

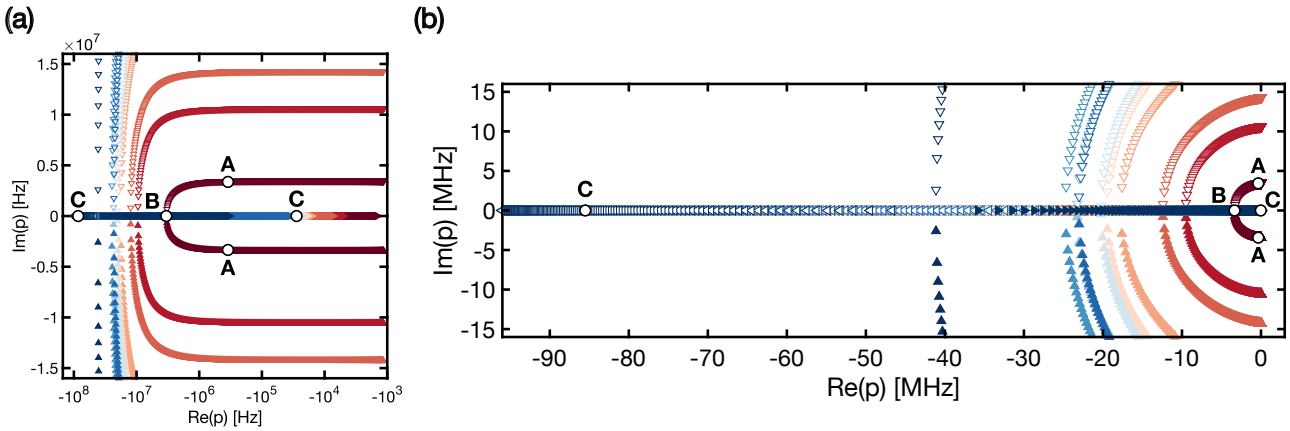

Fig. 7. Root locus of the circuit poles for the linear system matrix of Fig. 6, $\kappa_{\mathbf{X}} = 27$, for increasing values of the parameter f, on (a) semilogarithmic scale and (b) linear scale. Each pole pair starts as a complex-conjugate pair for low values of f (A), moving towards the real axis as f increases. The optimal value of f from a solution time standpoint is that for which the dominant pole pair reaches the real axis (B). Further increasing f leads to pairs splitting into high- and low-frequency components (C).

pair, whose real part increases linearly with f, whereas the imaginary part is mostly independent of the same parameter. On the other hand, for $f \to +\infty$, $\mathbf{U_1} \simeq f\mathbf{I}$ and thus $q_2(x) \to 1, q_3(x) \to \tilde{q}_3(x)/f \to 0$ where $\tilde{q}_3(x) \neq 0$. By Taylor's expansion, two roots are then found as $\lambda_+ = -\tilde{q}_3(x)/f$ and $\lambda_- = -1+\tilde{q}_3(x)/f$. By increasing the parameter f, therefore, the poles get closer to the real axis, eventually splitting into an high- and low-frequency components.

C. Dependences on circuit parameters

From our theory, $\mathbf{F}$ (or its scalar counterpart f) and $\kappa_{\mathbf{X}}$ play a major role in the accuracy-bandwidth trade-off of the proposed circuit. For the sake of simplicity, we shall refer to the scalar case $\mathbf{F} = f\mathbf{I}$ (namely, linear regression or linear system solver without preconditioning), but the same discussion holds as long as $\sigma_1(\mathbf{F})$ and $\sigma_n(\mathbf{F})$ are appropriately considered in place of f.

Fig. 6 shows the calculated transient of evolution of v_β and v_ε for various f, namely $f = 0.025$ (a), $f = 0.25$ (b) and $f = 15.8$ (c). As f is increased the circuit dynamics moves from a second-order regime (Fig. 6a) with damped oscillations, to an overdamped first-order regime (Fig. 6c), with an intermediate optimally damped case (Fig. 6b) achieving minimum solution time, in accordance with the position of poles predicted by Eq. (14). Fig. 7 shows the positions of the poles for increasing f, marking the three conditions in Fig. 6 as A, B, and C respectively. Fig. 8a shows the real part of the poles as a function of f: in both regimes, the real part of the poles depends linearly on f as predicted by Eq. 14. The optimal damping is obtained when the dominant pole pair reaches the real axis, i.e. for $q_2(x) = 2\sqrt{q_3(x)}$. Fig. 6c shows the real part of the pole as a function of $\kappa_{\mathbf{X}}$, which controls the $q_3(x)$ term. In the complex conjugate regime, i.e. for $q_3(x) > q_2(x)^2/4$, $q_3(x)$ controls the imaginary part, hence the ringing component, with no impact on the solution time. In the real regime, $q_3(x)$ controls the spread of singularities between the amplifiers' gain-bandwidth product (GBWP) and the dominant pole ($\omega_0 = 1/\tau_0$) thus dictating the position of the dominant pair (Fig. 8b). A perfectly conditioned problem, $\kappa_{\mathbf{X}} = 1, \mathbf{X} = \mathbf{I}$, corresponds to a circuit configuration with n decoupled, unity-gain loops, which therefore achieves the

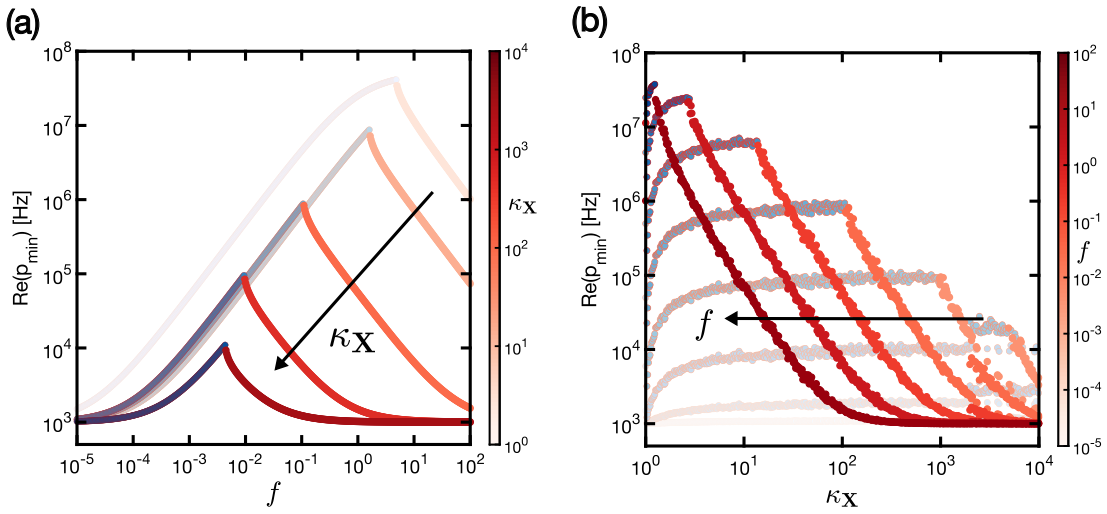

Fig. 8. Dependence of dominant pole on f and $\kappa_{\mathbf{X}}$. (a) Real part of the dominant pole as a function of f, for different values of $\kappa_{\mathbf{X}}$. The real part initially increases with f due to complex-conjugate pairs moving along their respective trajectories towards higher real parts, and then decreases as a consequence of pole splitting when the dominating pair becomes real. The critical f value depends on the condition number of the matrix $\kappa_{\mathbf{X}}$. (b) Real part of the dominant pole as a function of $\kappa_{\mathbf{X}}$, for different values of f. Blue-filled points represent complex-conjugate pairs, whereas red-filled points denote a real pole. The real part decreases linearly with $\kappa_{\mathbf{X}}$ after a critical point which varies with f. For high values of $\kappa_{\mathbf{X}}$, the dominant pole saturates to the OA's internal pole.

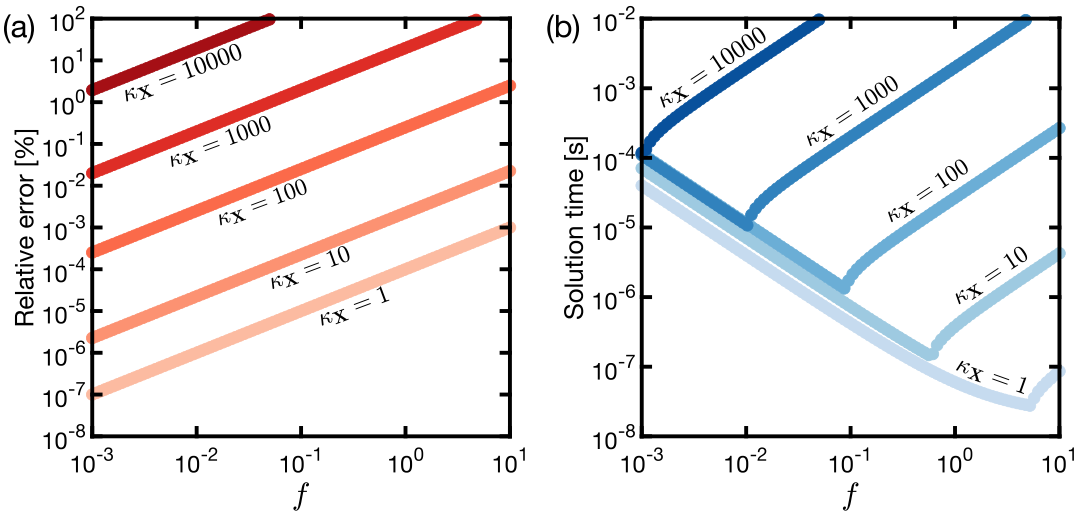

Fig. 9. Final performance of the circuit in operation as linear system solver without preconditioning. (a) Relative error with respect to exact solution as a function of the parameter f, showing an $\mathcal{O}(f\kappa_{\mathbf{X}}^2)$ dependence. (b) Impact of f on the solution time, showing a $\mathcal{O}(1/f)$ dependence up to the critical f, which varies with $\kappa_{\mathbf{X}}$, and a $\mathcal{O}(f\kappa_{\mathbf{X}}^2)$ dependence above. For small values of f, the circuit bandwidth is independent of the problem's condition number.

maximum possible bandwidth ($\alpha_0\omega_0$). On the other hand, as $\kappa_{\mathbf{X}}$ increases, $q_3(x)$ approaches singularity and one of the pole pairs quadratically approaches the dominant pole as one of its roots λ approaches 0.

From the accuracy standpoint, Eq. 11 provides the dependence on f and $\kappa_{\mathbf{X}}$. The quadratic dependence on $\kappa_{\mathbf{X}}$ can be intuitively explained by the presence of two crosspoint matrices $\mathbf{X}$ along a single loop, whereas the dependence on f can be explained by considering the nodal equation at the A_1 inputs, where the nonzero current drawn from the f conductance perturbs the otherwise exact solution of the regression operation.

Fig. 9 summarizes the impact of f and $\kappa_{\mathbf{X}}$ on the solution time and error for the cases of linear regression and linear system solver without preconditioning. For low f, in the complex/underdamped regime, the solution time is almost independent of $\kappa_{\mathbf{X}}$, and decreases with f only. On the other hand, ε_s increases with f and $\kappa_{\mathbf{X}}$.

IV. Impact of Parasitic Elements

FROM the error standpoint, the parasitic interconnect resistance leads to voltage, also known as IR, drops along

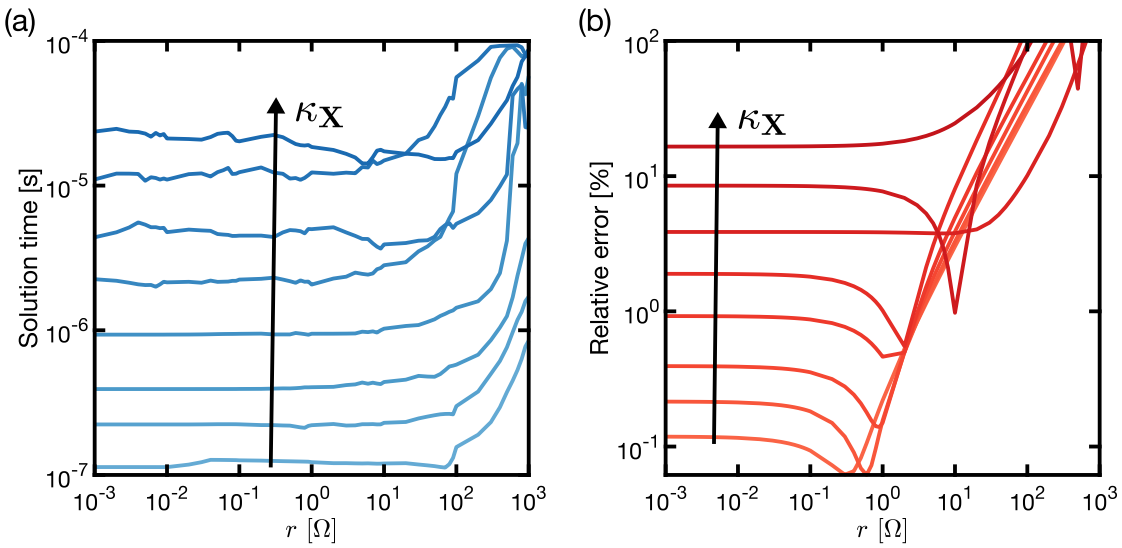

Fig. 10. Impact of IR drop on (a) solution time and (b) static error of the circuit. While the solution time is mostly unaffected by parasitic resistances, deviating from the ideal behaviour only for relatively large and untypical resistance values, the static error is more sensitive, showing an $\mathcal{O}(r)$ dependence for values falling inside the typical interconnect resistance range.

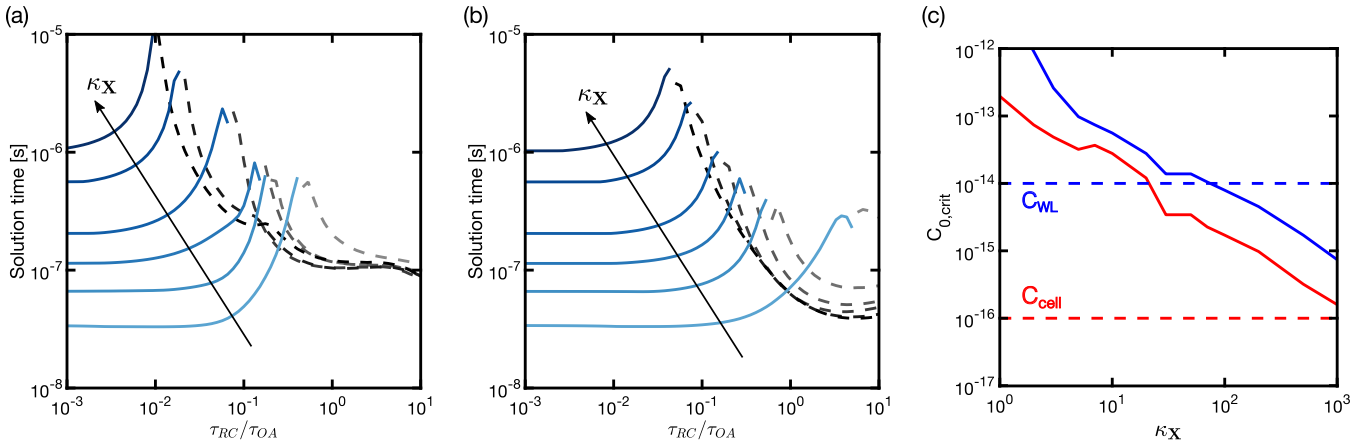

Fig. 11. (a) Impact of parasitic cell capacitances on the solution time. Each line is a median of 1000 simulations with parasitic capacitances extracted from a Gaussian distribution centered around $C_0 = \tau_{RC} G_0$ with $\sigma = C_0$. As the equivalent cell time constant approaches the closed-loop OA time constant, τ_{OA}, the solution time increases as the poles progressively move towards the origin (blue, solid lines). Further increasing C_0 leads to the dominant pole becoming unstable (grey, dashed lines). (b) Impact of parasitic line capacitances on the solution time. Each line is a median of 1000 simulations with parasitic capacitances extracted from a Gaussian distribution with the same parameters in (a). (c) Critical C_0 for cell (red) and line (blue) capacitance, defined as the mean capacitance around which the circuit becomes unstable. Dashed lines show typical values of $C_{cell} = 0.1\,\text{fF}$ (red) and $C_{WL} = 10\,\text{fF}$ (blue), highlighting the stronger constraint on $\kappa_{\mathbf{X}}$ given by C_{WL}.

the top and bottom electrode lines, which increases ε_s. Fig. 10 shows the impact of the average parasitic resistance r between two adjacent cells on the solution time and static error of the solution computed by the circuit for a problem of size $n = 10$, with $G_0 = 100\,\mu\text{S}$. Particularly, the solution time is mostly unaffected, deviating from the ideal behaviour only as r enters the $\text{k}\Omega$ range, which is typically not the case in most technologies [21]. On the other hand, the relative error shows a linear dependence for values of r greater than a characteristic value, which generally scales with n [22], comparable with the typical interconnect value of $1\,\Omega$. Nonetheless, dedicated compensation techniques [23] and parasitic-aware programming schemes may help alleviating this issue.

The error due to conductance non-linearity may be reduced by adopting relatively low resistance values [8], at the cost of increased energy consumption and IR drop. Finally, additional non-idealities in the operational amplifiers such as offset voltages may cause output errors increasing as $\kappa_{\mathbf{X}}^2$, or even circuit saturation for large $\kappa_{\mathbf{X}}$, due to the large swing of the outputs even in presence of small inputs. Dedicated offset-canceling circuits may mitigate the saturation issues at the cost of an $\mathcal{O}(n)$ area overhead.

From the speed standpoint, the parasitic cell and word-line capacitance may slow down the circuit for increasing n. On the other hand, as long as the circuit remains stable, no additional contribution to the static error is expected from parasitic capacitances. Figs. 11a-b show the impact of C_{cell} and C_{WL} on the circuit solution time. As long as the time constant of the parasitic capacitances $\tau_{RC} = C_0/G_0$ is much faster than those of the singularities attributable to the OA's finite

bandwidth, the solution time is unaffected by the presence of parasitics. On the other hand, as the capacitance increases, a coupling between the OA poles and the parasitic poles ensues, reducing the circuit's bandwidth as τ_{RC} becomes comparable with the closed-loop OA time constant, $\tau_{OA} = 1/GBWP$. Eventually, poles move across the origin into the Right Half-Plane, rendering the circuit unstable. The stability boundary shows a linear dependence on the parameter $\kappa_{\mathbf{X}}$, as shown in Fig. 11c. With typical values of $C_{cell} = 0.1\,\mathrm{fF}$ [22], $C_{WL} = 1 \sim 10\ \mathrm{fF/\mu m}$ [21], the major threat to bandwidth and stability comes from the word-line parasitic capacitances, including both metal line and OA input capacitance.

Finally, the non-zero output resistance of the operational amplifiers may reduce the useful bandwidth of the circuit due to the reduced strength of the local feedback caused by the additional voltage drop on r_{out}. For instance, the bandwidth is reduced by a factor 2 when the overall column conductance becomes comparable to the output resistance of the amplifier, leading to $n \leq 100$ for $r_{out} = 50\,\Omega$. Specialized topologies [24], [25] can be employed to reduce the output impedance and its impact on speed.

V. Conclusion

We present a novel circuit capable of performing various types of linear regression over a given dataset. We derive closed-form equations for the static error and the singularities, and assess the circuit performance in terms of solution bandwidth and error. We discuss the impact of additional nonidealities such as interconnect parasitic resistance and capacitance, operational amplifier offset voltage and output resistance. The circuit provides a universal primitive for linear algebra operations, paving the way for large-scale integration in a dedicated computing architecture exploiting the IMC paradigm.

Acknowledgment

This work was supported in part by the European Research Council (grant ERC-PoC-2018-842472-CIRCUS, grant ERC-CoG-2014-648635-RESCUE).

References

[1] M. A. Zidan, J. P. Strachan *et al.*, *Nature Electronics*, vol. 1, no. 1, pp. 22–29, Jan. 2018. doi: 10.1038/s41928-017-0006-8

[2] N. P. Jouppi, C. Young *et al.*, *ACM SIGARCH Computer Architecture News*, vol. 45, no. 2, pp. 1–12, Sep. 2017. doi: 10.1145/3140659.3080246

[3] D. Ielmini and H.-S. P. Wong, *Nature Electronics*, vol. 1, no. 6, pp. 333–343, Jun. 2018. doi: 10.1038/s41928-018-0092-2

[4] J. J. Yang, D. B. Strukov *et al.*, *Nature Nanotechnology*, vol. 8, no. 1, pp. 13–24, Jan. 2013. doi: 10.1038/nnano.2012.240

[5] D. Ielmini and G. Pedretti, *Advanced Intelligent Systems*, vol. 2, no. 7, p. 2000040, Jul. 2020. doi: 10.1002/aisy.202000040

[6] M. Hu, C. E. Graves *et al.*, *Advanced Materials*, vol. 30, no. 9, p. 1705914, Mar. 2018. doi: 10.1002/adma.201705914

[7] C. Li, D. Belkin *et al.*, *Nature Communications*, vol. 9, no. 1, p. 2385, Dec. 2018. doi: 10.1038/s41467-018-04484-2

[8] C. Li, M. Hu *et al.*, *Nature Electronics*, vol. 1, no. 1, pp. 52–59, Jan. 2018. doi: 10.1038/s41928-017-0002-z

[9] F. Cai, S. Kumar *et al.*, *Nature Electronics*, vol. 3, no. 7, pp. 409–418, Jul. 2020. doi: 10.1038/s41928-020-0436-6

[10] G. Pedretti, P. Mannocci *et al.*, *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 6, no. 1, pp. 89–97, Jun. 2020. doi: 10.1109/JXCDC.2020.2992691

[11] M. A. Zidan, Y. Jeong *et al.*, *Nature Electronics*, vol. 1, no. 7, pp. 411–420, Jul. 2018. doi: 10.1038/s41928-018-0100-6

[12] M. Le Gallo, A. Sebastian *et al.*, *Nature Electronics*, vol. 1, no. 4, pp. 246–253, Apr. 2018. doi: 10.1038/s41928-018-0054-8

[13] M. R. Williams, *A history of computing technology*, ser. Prentice-Hall series in computational mathematics. Englewood Cliffs, N.J: Prentice-Hall, 1985.

[14] J. S. Small, *Analogue Alternative: the Electronic Analogue Computer in Britain and the USA, 1930-1975*. London: Routledge, 2013, oCLC: 864391135. http://www.myilibrary.com?id=500770

[15] C. Mead, *Analog VLSI and neural systems*, ser. Computation and neural systems series. Reading, Mass: Addison-Wesley, 1989.

[16] Z. Sun, G. Pedretti *et al.*, *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4123–4128, Mar. 2019. doi: 10.1073/pnas.1815682116

[17] Z. Sun, E. Ambrosi *et al.*, *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1466–1470, Apr. 2020. doi: 10.1109/TED.2020.2966908

[18] Z. Sun, G. Pedretti *et al.*, *Science Advances*, vol. 6, no. 5, p. eaay2378, Jan. 2020. doi: 10.1126/sciadv.aay2378

[19] R. A. Horn and C. R. Johnson, *Matrix analysis*, 2nd ed. Cambridge ; New York: Cambridge University Press, 2012.

[20] F. Tisseur and K. Meerbergen, *SIAM Review*, vol. 43, no. 2, pp. 235–286, Jan. 2001. doi: 10.1137/S0036144500381988

[21] "International Technology Roadmap for Semiconductors (ITRS)," 2013. www.itrs2.net/2013-itrs.html

[22] M. E. Fouda, A. M. Eltawil *et al.*, *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 1, pp. 270–282, Jan. 2018. doi: 10.1109/TCSI.2017.2714101

[23] S. N. Truong, *Micromachines*, vol. 10, no. 10, p. 671, Oct. 2019. doi: 10.3390/mi10100671

[24] I. Mucha, *Electronics Letters*, vol. 28, no. 22, pp. 2071–2072, Oct. 1992. doi: 10.1049/el:19921327

[25] H. Barthelemy, *Electronics Letters*, vol. 33, no. 20, pp. 1662–1664, Sep. 1997. doi: 10.1049/el:19971170