



M Ű E G Y E T E M 1 7 8 2

SZAKDOLGOZAT

SZTOCHASZTIKUS JELENSÉGEK REZISZTÍV KAPCSOLÓ MEMÓRIÁKBAN

Fehérvári János Gergő
Természettudományi Kar

Témavezető: **Török Tímea Nóra**
doktorandusz
BME Fizikai Intézet
Fizika Tanszék

Budapesti Műszaki és Gazdaságtudományi Egyetem
2021. május 21.

Eredeti Szakdolgozattéma-kiírás

Sztochasztikus jelenségek rezisztív kapcsoló memóriákban

Stochastic phenomena in resistive switching memories

A rezisztív kapcsoló memóriák (memrisztorok) olyan eszközök, amelyek ellenállása programozható, azaz ellenállásuk tetszőlegesen változtatható az eszköz elektródáira kapcsolt elektromos jelekkel [1]. A memrisztorok ellenállása ezáltal információt hordoz, ráadásul az eszközök nem-illékony memóriaként hasznosíthatóak: az ellenállás kis jelszinten mérhető annak megváltozása nélkül. A rezisztív kapcsolás során az elektródák között található, kezdetben szigetelő vagy félvezető anyagban lokális anyagszerkezet-változás megy végbe, és ennek eredményeképp megváltozik az eszköz ellenállása. Az ellenállás változását rendszerint egy önszerveződő folyamatban kialakuló, fémes vezetési tulajdonságú filamentum okozza, kapcsolatot létesítve az eszköz elektródái között.

A rezisztív kapcsoló memóriák közé tartozó fázisváltó memóriákban az ellenállás-változás az aktív közeg kristályos/amorf állapotához köthető [2]. A sztochasztikus működésű SiO_x alapú rezisztív kapcsoló memóriák gazdag kapcsolási időskáláinak [3] részletes feltárása elősegíti a rezisztív kapcsolás hátterében álló mikroszkopikus fizikai folyamatok megértését. A szakdolgozatot készítő hallgató feladata grafén/ SiO_x /grafén fázisváltó memóriák kísérleti vizsgálata, ami magában foglalja az eszközök sztochasztikus viselkedésének tanulmányozására irányuló mérések tervezését, megvalósítását, illetve az adatfeldolgozást és az eredmények értelmezését. Ezen túlmenően célunk olyan probabilisztikus elvű számítási algoritmus keresése, ami hasznosítja a memrisztorok eredendően sztochasztikus természetét [4].

Önállósági nyilatkozat

Alulírott Fehérvári János Gergő a Budapesti Műszaki és Gazdaságtudományi Egyetem fizika BSc szakos hallgatója kijelentem, hogy ezt a szakdolgozatot meg nem engedett segédeszközök nélkül, önállóan, a témavezető irányításával készítettem, és csak a megadott forrásokat használtam fel. Minden olyan részt, melyet szó szerint, vagy azonos értelemben, de átfogalmazva más forrásból vettem, a forrás megadásával jelöltem.

Budapest, 2021. május 21.

Fehérvári J. Gergő

Digitális és nyomtatott dolgozatok egyezése

Alulírott Fehérvári János Gergő a Budapesti Műszaki és Gazdaságtudományi Egyetem fizika BSc szakos hallgatója kijelentem, hogy jelen szakdolgozat digitális és nyomtatott kiadása, melyeket az egyetem felé továbbítottam, mindenben megegyezik.

Budapest, 2021. május 21.

Fehérvári J. Gergő

Tartalomjegyzék

1. Bevezetés	1
2. SiO_x memrisztorok és vizsgálatuk	3
2.1. Fázisváltó memóriák alaptulajdonságai	3
2.2. Grafén elektródás SiO _x memrisztorok	5
2.2.1. Kísérleti technikák	5
2.2.2. Alaptulajdonságok	8
2.3. Kísérleti eredmények grafén elektródás SiO _x memrisztorokon	9
2.3.1. Írási idő eloszlása	9
2.3.2. Többosztintű programozhatóság	11
2.3.3. Kereszteffektusok	11
3. Komplex problémák megoldása memrisztorhálózatokkal	14
3.1. Neurális hálózatok	14
3.1.1. Crossbar struktúra	14
3.1.2. Hardveres implementálás	16
3.2. Valószínűségi optimalizálás Hopfield-hálózatokkal	17
3.3. Kísérleti megvalósítás	18
3.3.1. A sztochaszticitás szerepe	19
3.3.2. A sztochaszticitás forrása	20
3.3.3. A hálózat teljesítménye	23
3.4. Hopfield-hálózat szimulációja	24
4. Összefoglalás és kitekintés	27
5. Köszönetnyilvánítás	28
A. Hopfield hálózat energia-minimalizációs tulajdonsága	29
B. Max-cut probléma mint optimalizációs feladat	31

1. Bevezetés

A Moore-törvény végéhez közeledve az integrált áramkörök gyártásában az információátvitel effektív mérete lassan eléri az atomi méretskálát. Ezzel párhuzamosan, a mesterséges intelligencia és a gépi tanulási algoritmusok fejlődésével, új számítástechnikai paradigmák szükségesek napjaink növekvő adatsokaságának feldolgozásához. Kisméretű és energiahatékony számítógépes memóriaegységek és hardveres gyorsítók alapegységei lehetnek a memrisztorok (memóriával rendelkező rezisztorok). Ezek olyan fizikai rendszerek, amelyek külső elektromos jelek hatására képesek vezetőképességüket változtatni, a hatás megszűntével pedig megőrzik felvett állapotukat. A memrisztorok egy alcsoportját képezik a fázisváltó memóriák, amelyekben külső feszültségimpulzusok hatására fázisátalakulás megy végbe. A feszültségjelekkel a magas vezetőképességű kristályos és alacsony vezetőképességű amorf állapotok között tudunk váltani. Ezt a folyamatot alacsony-magas átmenet esetén írásnak, magas-alacsony átmenet esetén pedig törlésnek nevezzük. Kis amplitúdójú feszültségjelekkel az egyes állapotok vezetőképessége azok megváltoztatása nélkül kiolvasható, így ezek az eszközök, nem illékony memóriaegységként alkalmazhatók.

A fenti előnyös tulajdonságokon túl a memrisztorok rendszerint többszintű programozhatóságot mutatnak. Ez azt jelenti, hogy a gerjesztő feszültségimpulzus paramétereinek a hangolásával az eszköz magas vezetőképességű állapota több lépésben is változtatható. A memrisztorok úgynevezett crossbar sémába való rendezésével (lásd 3.1 szakasz) tehát kísérletileg is képesek lehetünk egy elemi fizikai törvényeken alapuló hardvert építeni a mátrix-vektor szorzások gyors, nagyfokú párhuzamosított elvégzésére. Ez teret ad a matematikájukban jórészt mátrix-vektor szorzásokra visszavezethető neurális hálózatok fizikai implementálásának is, felgyorsítva működésüket. A szoftveres neurális hálózatok esetében kifejezetten számításigényes művelet a neurális rétegek közötti átmenet szinaptikus súlyoknak megfelelő kiértékelése. A memrisztorokból felépített hálózatok azonban ezen művelet egyetlen lépésben történő, hardveres elvégzését teszik lehetővé.

Jelen dolgozatban *grafén elektródás szilícium-oxid memrisztorok* vizsgálatával foglalkozom. Ezekben az eszközökben a kapcsolható aktív tartomány egy termikusan növesztett SiO_x vékonyrétegben alakítható ki. Az oxidréteg felszínére transzferált grafénrétegben kialakított nanovezetékek kontrollált elvékonyítása és elszakítása után elektródákon keresztül tudunk elektromosan hozzáférni az eszközhez. A SiO_x memrisztor sajátossága, hogy az írási- és törlési folyamatok jellemző időskálái mellett megfigyelhető benne az ún. *holtidő* is [3, 5]. Ez egy átmeneti időtartam a törlést követően, amíg nem lehet ismételt írást végrehajtani. Különlegessége, hogy lehetővé teszi az amorf állapot vezetőképességéhez való hozzáférést zérus feszültségen.

A három karakterisztikus időskálának köszönhetően a SiO_x alapú fázisváltó memóriák komplex dinamikai viselkedést mutatnak. TDK munkám keretében korábban ezt a komplex viselkedést igyekeztem legalább részben feltárni, behatóan vizsgálva az írási idők eloszlását és hangolhatóságát, az írási- és törlési folyamatok közötti kereszteffektusokat, illetve kimutatva a holtidő alatt végbemenő folytonos vezetőképesség-relaxációt. Mintegy kísérleti munkám folytatásaként szakdolgozatomban új, korábbi eredményeimet alátámasztó mérést és az írási idők eloszlására magyarázatot adó elméleti modellt prezentálok. Ezen túlmenően bemutatom a memrisztorok alkalmazhatóságát vizsgáló irodalomkutatás [4] és szimuláció eredményeit. Mindkét témának külön fejezetet szenteltem,

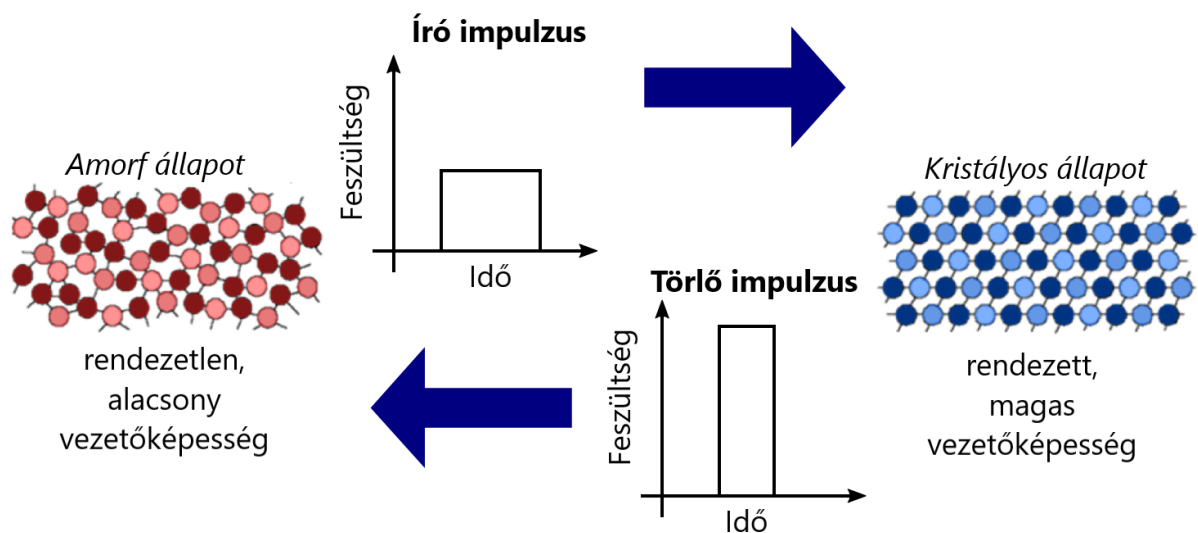
melyek mindig a releváns ismeretek és irodalmak rövid ismertetésével kezdődnek. A 2. fejezet a SiO_x memrisztorok időskáláinak behatóbb kísérleti vizsgálatával foglalkozik, a 3. fejezetben pedig az eszközök Hopfield-hálózatokban való alkalmazhatóságát tárom fel szimulációs módszerekkel.

2. SiO_x memrisztorok és vizsgálatuk

Ebben a fejezetben először bemutatom a fázisváltó memóriák, majd specifikusan a grafén elektródás SiO_x memrisztorok ismert jellemzőit. Ezt követik saját TDK munkám legfőbb eredményeinek összefoglalása és a szakdolgozathoz készült mérési eredmények ismertetése és értelmezése.

2.1. Fázisváltó memóriák alaptulajdonságai

A memória rezisztorok, vagy röviden memrisztorok olyan, általában fémrétegek közé szendvicselt szigetelőből álló nanoszerkezetek, amelyek vezetőképessége változtatható és legalább kettő jól elkülönülő értéket vehet fel. Az így megkülönböztethető állapotokat alacsony vezetőképességű *OFF* és magas vezetőképességű *ON állapotoknak* nevezzük. Közöttük meghatározott külső behatásra - például az eszköz elektródáira kapcsolt elektromos impulzusok hatására - váltás idézhető elő. Ez a rezisztív kapcsolás folyamata, ami lehetővé teszi például az eszköz vezetőképességében kódolt adattárolást. A memrisztorok fontos tulajdonsága, hogy nem-illékonyak, vagyis külső erőforrás nélkül is megőrzik az információt.



2.1. ábra. A fázisváltó memóriák alacsony és magas vezetőképességű állapotainak sematikus rajza [2]. Látható az amorf és kristályos állapot illusztrációja, illetve az átmenetet előidéző gerjesztő feszültségimpulzusok jellemző alakja.

A vezetőképesség változásának mechanizmusa szerint több elrendezésben is alkotunk rezisztív kapcsoló eszközöket. Egy lehetőség a fázisváltó (*PCM, phase-change memory*) memória gyártása, melyben a fázisváltozást külső feszültségimpulzusok idézik elő. Az átalakulást okozó fizikai folyamat elektromos tér- vagy hőmérséklet-indukálta átmenet egy anyag magas vezetőképességű kristályos és alacsony vezetőképességű amorf állapotai között, ahogy a 2.1. ábrán is látható [2]. Az eszköznek a fázisváltásban résztvevő térfogatát aktív tartománynak nevezzük. Ennek amorfizációja egy gyors lefolyású folyamat, ami akkor következik be, amikor az eszközre kapcsolt feszültség hajtotta árammal járó Joule-hő eléri a kristályos állapot olvadáspontját. Ekkor a rendezett struktúra

megolvad, hirtelen megszűnik a fűtőhatás és az atomok egy rendezetlen állapotban ragadnak. Az atomi mozgékonyság gyors lecsökkenése tulajdonképpen az amorf állapotba való befagyást eredményez. Ezt a vezetőképesség tekintetében magas-alacsony átmenetet *törlésnek* szokás nevezni. Ennek megfelelően az aktív tartomány kristályosodására, az alacsony-magas vezetőképességű átmenetre a konvenció szerint *írás*ként hivatkozunk. Az írás szintén egy gyors folyamat, melyben az elektromos tér rendező hatása érvényesül, ez idézi elő az atomok rendeződését.

Az írási és törlési folyamatok időbeli viselkedésének leírására bevezethetjük az $\tilde{t}_{\text{írás}}$ *írási* és $\tilde{t}_{\text{törlés}}$ *törlési idők* fogalmát. Ez mindkét esetben a szükséges feszültségimpulzus megjelenését követően eltelt idő, ami a fázisátalakulás megkezdődéséhez szükséges. Az átalakulás lefolyása a várakozási időknél nagyságrendekkel kisebb, sőt általában az írás és törlés is eltérő időskálákon zajlanak. Érdekes következtetésre juthatunk, ha az írási folyamatot magképződéssel modellezzük [6]. Ebben a képben az időigényes lépés az első kristályos mag kialakulása az amorf tartományban, amit a tartomány teljes és gyors átkristályosodása követ. A magképződést (*nukleációt*) a

$$\tau = \tau_0 \cdot \exp\left(\frac{W}{k_B T}\right)$$

időskála jellemzi, ahol τ_0 konstans, W a nukleációs energia, k_B a Boltzmann-állandó, T pedig a hőmérséklet. A mag kialakulásának valószínűsége ekkor egyparaméteres Poisson-folyamattal írható le, vagyis az írási idő egy τ paraméterű exponenciális eloszlást követ

$$f_{\tilde{t}_{\text{írás}}} \propto \exp\left(-\frac{t}{\tau}\right).$$

Az írási idők eloszlására tehát levezethető egy elméleti eloszlás, ennél is érdekfeszítőbb azonban, hogy ha feltesszük, hogy a nukleációs energia

$$W \sim \mathcal{N}(W_0, \sigma_W^2)$$

Gauss-eloszlást követve ~ 0.1 eV nagyságrendben fluktuál időben, akkor belátható, hogy ez fluktuációt idéz elő az írási idő eloszlásában is. A τ nukleációs időn keresztüli csatlás révén tehát az írási idő eloszlása is módosul. A $\frac{\sigma_W}{W_0}$ relatív szórásra levezethető egy kritikus értéktartomány, ami felett az eloszlás eredő képe lognormális lesz, ezalatt pedig a fluktuáció és az eloszlás torzulása elhanyagolható. Ekkor a fentiekben megadott exponenciális eloszlással jól leírhatóak az írási idők. Kísérleti munkámban nagyszámú eloszlás rögzítésével meg tudtam mutatni, hogy néhány mérés időskáláján a nukleációs energia fluktuációja kellően kicsi, így az eloszlás valóban exponenciálisnak mérhető, hosszabb időskálákon azonban a fluktuáció felerősödik és valóban lognormális eloszlást figyelhetünk meg. Erről részletesebben a 2.3.1. alfejezetben számolok be.

A klasszikus elektróda – fázisváltó anyag – elektróda elrendezésű memrisztorokban a feszültségimpulzusok az elektródákon keresztül hőt közvetítenek. Ez inhomogén hőmérséklet-eloszláshoz vezet, ami determinálja, hogy kis amplitúdójú/rövid író feszültségimpulzusok mely tartományokat kristályosíthatják át könnyen. Így általában finomabb skálán is szabályozható a kristályos állapotban lévő aktív tartomány vezetőképessége. A körültekintően paraméterezett feszültségjelek tehát az egyes ellenállás-állapotok többszintű programozhatóságát (finomhangolhatóságát) is lehetővé teszik.

A fázisváltó rendszerek meglehetősen komplex viselkedést mutatnak, melyek *termikus*, *elektromos* és *szerkezet-dinamikai* effektusokra és ezek kölcsönhatására vezethetőek vissza. A rezisztív kapcsolás folyamata, legyen szó kristályosodásról vagy amorfizációról, erősen függ az elektromos tértől és a hőmérséklettől. Ennek végeredménye egy a karakterisztikus időskálák tekintetében gazdag dinamikai tulajdonságokat mutató eszköz, melynek belső működése a kapcsolási folyamatokban fellelhető sztochaszticitáshoz és különféle relaxációs jelenségekhez vezet.

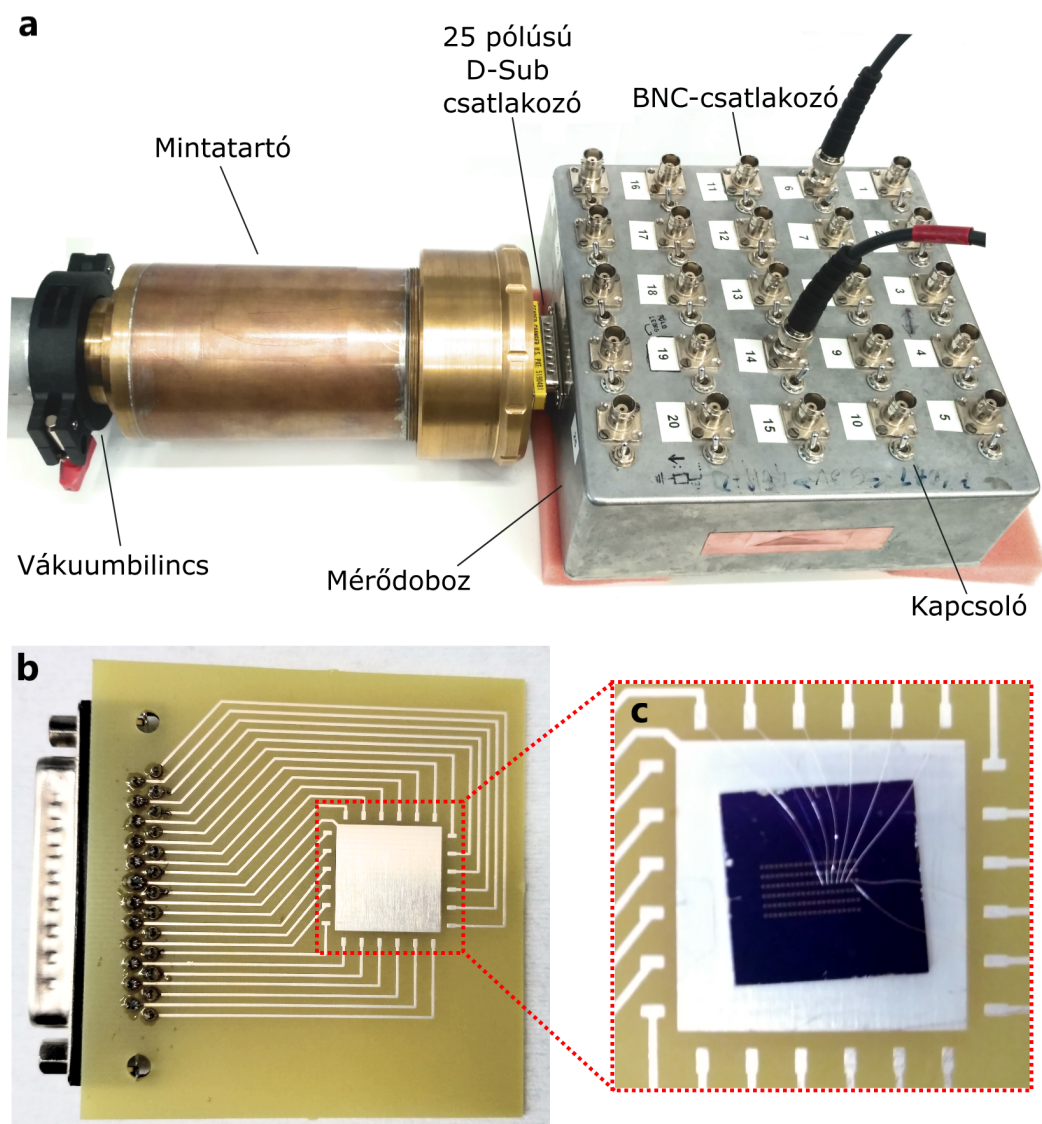
2.2. Grafén elektródás SiO_x memrisztorok

A SiO_x alapú fázisváltó memóriákat közelmúltban a BME Atomi és Molekuláris Elektronika kutatócsoportjában már vizsgálták, ezek a kapcsolási dinamika és az ún. holtidő feltérképezésére irányultak, kihasználva nm-es aktív tartomány létesítésének képességét [3, 5]. Pósa László doktori munkájából ismertem meg a legfontosabb tulajdonságokat és a kutatócsoport munkájához csatlakozva az eszközök dinamikai tulajdonságainak kísérleti vizsgálatával foglalkoztam [7]. A mérési elrendezés és a minták gyártásának bemutatását követően a SiO_x memrisztorok alapvető működését tehát saját feltáró méréseim eredményein keresztül, de a doktori munka vezérfonalát követve igyekszem prezentálni.

2.2.1. Kísérleti technikák

A vizsgálataimhoz szükséges mintákat Santa Botond és Pósa László készítették. Ebben a folyamatban először dópolt szilícium lapka csókályhában való oxidációjával felületi szilícium-oxid réteget hoznak létre. Ezután kémiai gőzfázisú leválasztással rézfelületre grafént választanak le, amit a szilícium lapkára transzferálnak. A grafénréteget polimerrel bevonva pásztázó elektronmikroszkóppal rajzolják meg az első maszkot. Oldószerrel eltávolítva a kijelölt tartományokat, felpárolgatásra kerülnek az elektromos kontaktust biztosító aranyelektródák. A polimer réteg teljes eltávolítása, majd újbóli polimerreteg felvitelét követően kerül fel, ismét pásztázó alagútmikroszkóppal megrajzolva, a grafén marási maszk. Ezt argon-oxigén plazmás marás követi, mely során lekerül a felületről a felesleges grafén, majd utolsó lépésként teljes acetonos depolimerizáció következik. Így rendelkezésünkre állnak az arany elektródákon keresztül hozzáférhető grafén nanovezetékek, alattuk pedig jelen van a szilícium hordozó felső, oxidált rétege. Munkám során az eszközöket ebben az állapotban vettem át.

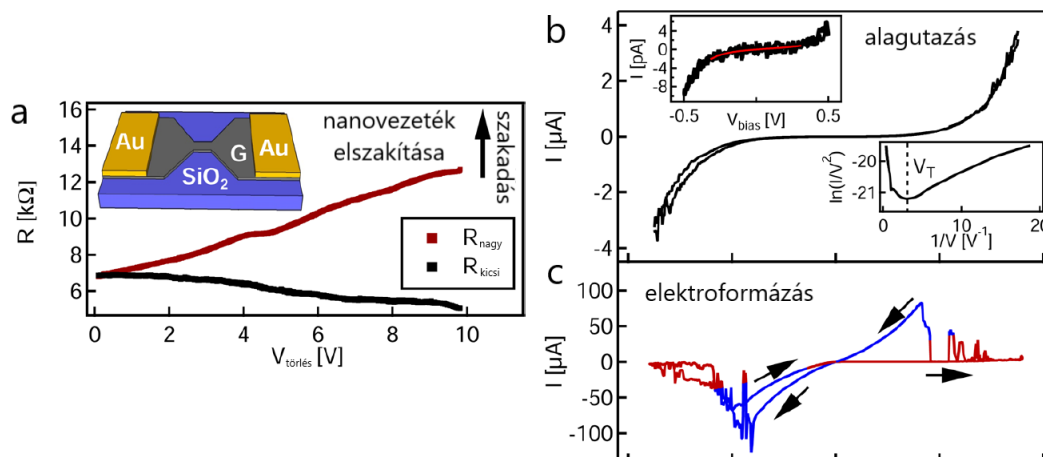
Ahhoz, hogy valóban memrisztorról beszélhessünk, az eszközöket elő kell készítenünk. Az első, hogy a chipet egy biztonságos, elektromos hozzáférést biztosító tartón rögzítjük. A 2.2.b ábrán látható a NYÁK, melynek központi fém felületére ezüsttartalmú ragasztóval rögzíthetjük a chipet. A következő a mikrohuzalkötések kialakítása, amivel az elektródákat a NYÁK felületi vezetékeihez rögzíthetjük. A bondolást, melynek végeredménye a 2.2.c ábráján látható rendszerint jómagam hajtottam végre. Egy bondolási folyamat során 5-6 elektródapárt lehet a NYÁK vezetékeihez huzalozni, ezt követően a kiválasztott mintákat a D-Sub csatlakozó megfelelő bemenetein keresztül címezhetjük. Az így előkészített minták ezután elektromos kezelést igényelnek. Először egy kontrollált folyamatban el kell vékonyítanunk, majd szakítanunk a grafén vezetékeket - így kaphatunk grafén elektródapárt - majd egy újabb kontrollált folyamatban preparálni kell az aktív tartományt - előkészítve azt az ismételt kapcsolásokhoz [5].



2.2. ábra. A mérőberendezés és a mintatartó a chipekkel. **a)** A vákuumtömör mintatartó; a benne lévő nyáklap D-Sub csatlakozókon keresztül érintkezik a kapcsolódobozzal. Ezen BNC-csatlakozókkal címezhetjük meg az egyes mintákat. **b)** A Pósa László tervei alapján készült NYÁK, melyen az egyes felszíni vezeték forrasztás útján köthetőek össze a csatlakozó kimeneteivel. A bekeretezett (.....) vezető lapkára Ag-tartalmú ragasztóval rögzítjük a chipet. **c)** A memrisztormátrixot tartalmazó, felragasztott chip. Az egyes elektródapárok a vezetékhez mikrohuzalkötések által csatlakoznak.

Ehhez természetesen a teljes mérőberendezésre szükség van, melynek legfontosabb elemei a 2.2.a ábrán láthatóak. A réz mintatartó kettős szereppel bír: elektromos árnyékolást biztosít és lehetővé teszi a nagyvákuum körülmények biztosítását, amit egy rotációs és egy turbomolekuláris szivattyú hoz létre két lépésben. A mintához a 25 pólusú D-Sub csatlakozón keresztül férhetünk hozzá, melynek minden portja egy-egy kétállású kapcsolóhoz van forrasztva. Itt válthatunk a nem használt minták megóvását célzó nagy ellenálláson keresztüli földelés és a BNC-csatlakozóra való kapcsolás között. A mikro-

huzalkötések készítése során kiválasztott vezeték egyértelműen meghatározzák a minta elérését szolgáló csatlakozókat. A mérésvezérlés egy National Instruments USB-6363 mérőkártyán keresztül történik, a meghajtó jel egy csatlakozóra jut, míg a másikon mért jel egy áramerősítőn keresztül jut vissza a kártyába. Szoftveresen pedig a Pósa László által C# nyelven implementált mérésvezérlő program felel az irányításért, melynek komplex jelalakok programozását célzó bővítésében magam is részt vettem. A mérőrendszer validációjával, illetve a felhasznált áramerősítők minősítésével TDK dolgozatomban részletesen foglalkoztam [7].

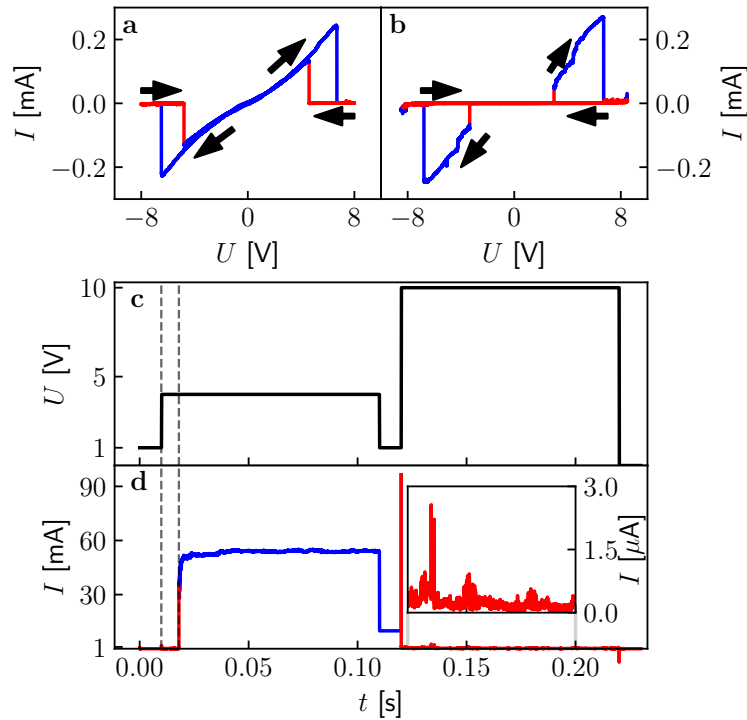


2.3. ábra. SiO_x memrisztor modellje és az elektromos kezelés lépései [3]. **a)** A grafén nanovezetékek elszakítása során a rövid, folyamatosan növekvő amplitúdójú jelek között, minden rövid formázó impulzus után egy alacsony feszültségű, hosszú olvasóimpulzus is kerül a mintára. Így a folyamat során a formázó (■) és olvasó (■) jelekkel mért ellenállások egyaránt megfigyelhetők. A magas feszültség a Joule-hő révén fűti a mintát, ami monton növekvő ellenállást eredményez, míg alacsony feszültségen monoton csökkenést látunk, ami a grafén tisztulását bizonyítja. **b)** A szakítást követően mért nemlineáris $I(V)$ karakterisztika, melynek lineáris tartományából a Simmons-modell segítségével megbecsülhetők a nanorés méretei. **c)** Kezdeti $I(V)$ mérések (elektroformázás), melyekkel néhány ismétlés után stabilizálható a kapcsolat [5].

Az előkészített mérőrendszerbe helyezve a mintát megkezdődhet a grafén nanovezetékek kontrollált szakítása. Az elektroformázás első lépése a grafén nanovezeték szabályozott elégetése. Ez $\sim 10 \mu\text{s}$ -os időskálájú, egyre növekvő amplitúdójú feszültségjelekkel történik, ahogy az a 2.3.a ábrán látható. A vezeték elszakítását követően alacsony vezetőképességű állapot tapasztalható - lásd 2.3.b ábráján - ami egy $\sim \text{nm}$ -es méretskálájú rés kialakulását jelzi. Ekkor megkezdődhet az eszköz szimmetrikus fűrészfog jellel való gerjesztésre adott áramválasz függvényének ismételt kimérése. Ezt $I(V)$ görbének nevezzük, és a preparáció ezen szakaszában nem csak karakterizálja az eszközt, hanem formázza is az aktív tartományt. A 2.3.c ábrán látható egy kezdeti instabil $I(V)$ görbe, mely után a kapcsolat stabilizálódik. Az így elektroformázott minták vezetőképessége ezután a grafén elektródák között található SiO_x rétegben kialakuló nanoskálájú aktív tartomány kristályos- vagy amorf állapotától függ, a memrisztorok tehát elő vannak készítve a működtetéshez.

2.2.2. Alaptulajdonságok

A 2.4.a ábrán látható egy előkészített eszközön mért tipikus $I(V)$ görbe. Itt a zérus feszültségű pontból pozitív feszültségtartományba elindulva nyilak jelölik a körüljárási irányt, a minta kezdetben ON állapotban található. Látható, hogy egy kritikus feszültségértéknél megkezdődik az átkapcsolás alacsony vezetőképességű OFF állapotba, az áram lecsökken. Az OFF állapotban maradva érjük el a maximális gerjesztő feszültséget és megkezdődik a gerjesztő jel amplitúdójának csökkenése. Egy újabb kritikus feszültségértéknél ugrásszerű vezetőképesség-növekedést tapasztalunk, az áram megnő, visszatérünk ON állapotba. Így lépünk át a negatív feszültségtartományba, ahol a karakterisztika a leírtaknak megfelelően ismétlődik. Ezt lassú áramválasznak nevezzük.



2.4. ábra. Az eszköz áramválaszai. A konvenció szerint kék szín kódolja az ON, piros szín pedig az OFF állapotot. **a)** Lassú $I(V)$ görbe, a szimmetrikus fűrészfogjellel történő gerjesztésre adott válasz. **b)** Gyors $I(V)$ görbe, itt az eszköz OFF állapotból indul és kétszeri kapcsolást követően a holtidő alatt éri el a zérusátmenetet, ami a gyors jelnek köszönhetően szintén OFF állapotban történik. **c)** Klasszikus feszültségimpulzus az írási idő mérésére, kisfeszültségű (1 V) ellenállásmérő, író (4 V) és törlő (10 V) szakaszokkal. **d)** A minta áramválasza a c) panelen látott gerjesztésre. A görbére jellemző írási idő határait szürke szaggatott vonal (---) jelöli.

Másfajta karakterisztikát látunk, ha a kezdetben OFF állapotban található mintára az előbbinél nagyobb frekvenciájú gerjesztő fűrészfogjelet adunk. Ekkor megfigyelhető a 2.4.b ábrán látható ún. gyors $I(V)$ görbe is. Itt kezdetben a növekvő gerjesztőfeszültség hatására megtörténik az írás, majd ezt követi egy törlés. OFF állapotban érjük el a maximális gerjesztést, azonban a csökkenő feszültség ellenére - látszólag meglepő módon - nem történik ismételt bekapcsolás, így OFF állapotban történik a zérusátmenet.

A karakterisztika a továbbiakban ismét szimmetrikusan ismétlődik. A görbét a holtidő segítségével értelmezhetjük: amíg ez nem telik le a kikapcsolást követően, addig az eszköz nem írható újra, így történhet zérusátmenet OFF állapotban.

A gerjesztéseket illetően az általános tapasztalat tehát az, hogy egy alacsonyabb (3–6 V-os) amplitúdójú feszültségimpulzussal az eszköz írható, míg egy magasabb (>8 V-os) amplitúdójú feszültségimpulzussal törölhető. Mindkét művelet ugrásszerű, de véges időt vesz igénybe a gerjesztés megkezdésétől, hogy megtörténjenek. Ezt a jelenséget mutatja a kapcsolási idők kimérésére alkalmazott mérőjel a 2.4.c és a vele mért áramválasz a 2.4.d ábrán. A gerjesztés kezdetétől számított időtartamot, ami a megfelelő állapotok közötti váltásig tart $\tilde{t}_{\text{írás}}$ írási időnek és $\tilde{t}_{\text{törlés}}$ törlési időnek nevezzük.

A 2.4.d ábrán mutatott mérésben kezdetben OFF állapotból indulunk, alacsony vezetőképeségű állapotban van az eszköz. Ezután kiadjuk rá az írófeszültséget és $\tilde{t}_{\text{írás}}$ időnek el kell telnie, mígnem ugrásszerűen megnő a mintán mért áram. Ez azt jelenti, hogy nagy vezetőképeségű ON állapotba érkezünk; a lezajlott folyamat pedig az írás jelensége. Az íróimpulzust követően egy zérus gerjesztésű szakaszt látunk, amit a nagyobb amplitúdójú törlőimpulzus követ. A ugrásszerű törlés kezdetéig itt $\tilde{t}_{\text{törlés}}$ időnek kell eltelnie. Figyeljük meg, hogy a törlési folyamat időskálája jóval gyorsabbnak mutatkozik az írásénál. Az eszköz végül ismét OFF állapotban marad. Mindkét kapcsolási idő sztochasztikus, azonos meghajtás mellett értékeik széles időtartományokat fedhetnek le, és várható értékük erősen függ a rezisztív kapcsolást indukáló feszültségimpulzus amplitúdójától.

2.3. Kísérleti eredmények grafén elektródás SiO_x memrisztorokon

Láttuk tehát, hogy a memrisztorok különböző állapotai közötti váltások dinamikus folyamatok, a kapcsolások megkezdődéséhez szükséges $\tilde{t}_{\text{írás}}$ és $\tilde{t}_{\text{törlés}}$ időparaméterek pedig több nagyságrendet átfogó valószínűségi változók. Ezek eloszlása feltehetőleg szoros kapcsolatban áll a kapcsolási dinamikával és így az eszköz működését leíró fizikával. A kutatócsoport korábbi vizsgálatai már felmérték az írási- és törlési idők feszültségfüggését [3, 5], ám az eloszlás behatóbb vizsgálatához nagyobb statisztikák felvétele volt szükséges [7].

2.3.1. Írási idő eloszlása

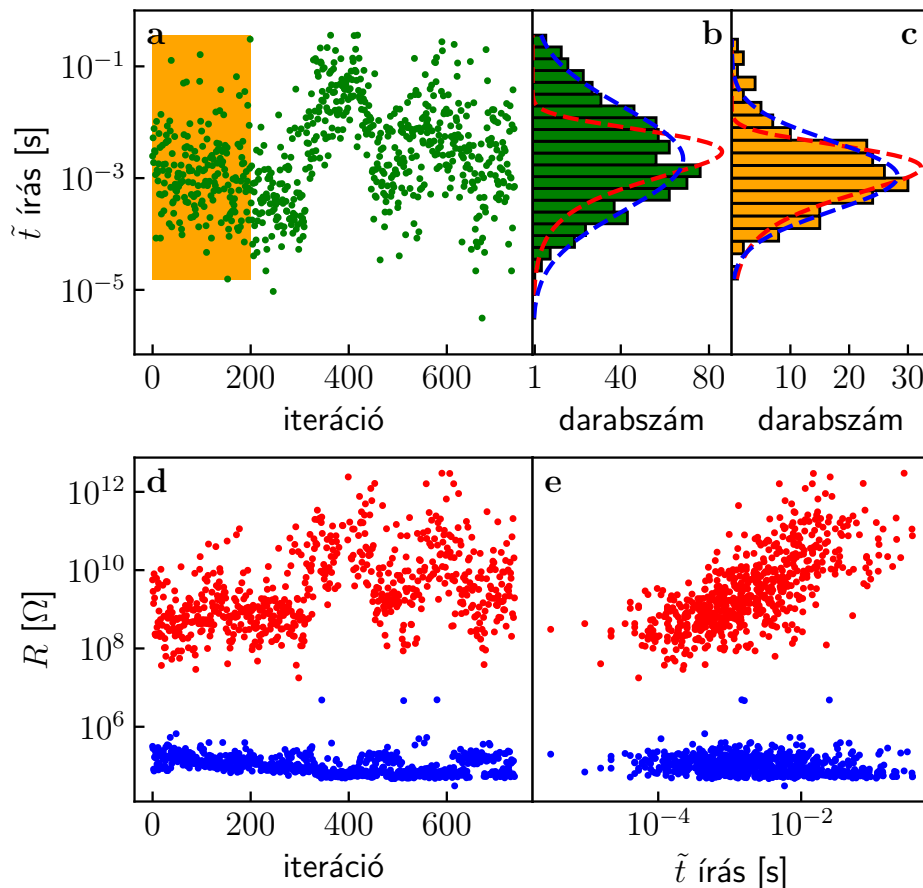
Ehhez TDK munkám során több adatsort is rögzítettem, különböző mintákon, eltérő meghajtójelek mellett. Megfigyelhető volt, hogy többszázas nagyságrendben végezve a méréseket az írási idő átlaga elhangolódik. Az adathalmazból logaritmikus hisztogramot készítettem. Ekkor az adatok blokkokba rendezése exponenciálisan növekvő szélességű lépésközökkel történik, így logaritmikus ábrázolás esetén kapunk azonos szélességű, ekvidisztáns tartományokat. A kapott görbéket

$$f_{\tilde{t}_{\text{írás}}}(t|\lambda) = \lambda \cdot e^{-\lambda t} \quad t > 0$$

exponenciális és

$$f_{\tilde{t}_{\text{írás}}}(t|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi t}} \cdot \exp\left(-\frac{(\ln(t) - \mu)^2}{2\sigma^2}\right) \quad t > 0$$

lognormális függvényekkel illesztettem, figyelembe véve a logaritmikus transzformációt. Egyértelműen megállapítható volt, hogy a homogén tartományokból készült hisztogramok exponenciális, a kiszélesedett, teljes mérési adatsorokból készült hisztogramok pedig lognormális függvénnyel írhatóak le.



2.5. ábra. Írási idő eloszlása. **a)** Az írási idő (●) időfejlődése hosszú mérési sorozatban. **b)** A kapcsolási idő eloszlása a teljes mérési adatsorra, ahol a lognormális függvény (---) esetén tapasztalunk precízebb illesztést. **c)** A kapcsolási idő eloszlása az első 200 pontra, itt az exponenciális függvénnyel (---) történő illesztés is kielégítően pontos. **d)** Az ON (●) és OFF (●) állapotú ellenállások a mérési index függvényében. Egyértelműen látható, hogy az OFF állapotú ellenállás átlagának elhangolódása jellegében leköveti az írási idő átlagának elhangolódását. **e)** Az ON (●) és OFF (●) állapotú ellenállások a mért írási idő függvényében. Szépen látható, hogy míg az ON állapot az írási időtől független, az OFF állapot vele spontán korrelációt mutat.

Azért, hogy a jelenséget még precízebben figyelhessem meg, újabb méréssorozat elvégzése volt szükséges. A járványhelyzetre való tekintettel a témavezetőm végezte a mérést, jómagam pedig a kiértékelést. A mérés során 953 egyszerű mérőjel (ld. 2.4.c ábra) került a mintára, rögzítve közben az áramválaszt. A görbékből számolt írási idők eredményei a 2.5.a ábrán láthatóak. Az írási idő az iterációk számának függvényében véletlenszerű drift-jelenséget produkál. A 2.5.b ábrán, az adatsorból készült hisztogramon megfigyelhető, hogy a korábbi eredményekkel egyezően az írási idők széles, nagyjából két

nagyságrendet átfogó tartományban szórnak. Szembetűnő ezenkívül, hogy a lognormális függvénnyel szépen illeszthető az adatsor, míg az illesztett exponenciális függvény mind kis, mind pedig nagy időskálákra jelentősen elüt a mért értékektől.

Érdemes azonban ismét megvizsgálnunk a mérési adatsor egy közel homogén tartományát - például az első 200 pontját - külön hisztogramon, ez tekinthető meg a 2.5.c ábrán. Az illesztett görbékéből kivehető, hogy az exponenciális eloszlás ezen a tartományon kielégítően írja le az adatsort. A végkövetkeztetésem tehát az, hogy a mérési eredmények a drift-jelenségek fellépése esetén lognormális eloszlásképet mutatnak, azonban elegendő számosságú homogén adatsorra a hisztogramok ugyanolyan jól, vagy akár jobban illeszthetők exponenciális sűrűségfüggvényekkel. A két eloszlás együttes megjelenése pedig alátámasztja, hogy a bekapcsolás során indukált kristályosodási folyamat történik, amelyre hatással lehet az eszköz OFF állapotában az amorf tartomány kapcsolási ciklusonként változó mérete is [7, 8].

Nem kerülheti el a figyelmünket továbbá a 2.5.d és a 2.5.e ábrák tartalma sem. Itt az ON és OFF állapotú ellenállások iterációnkénti időfejlődése, illetve az írási idővel vett korrelációjuk látható. Ebből kiderül, hogy az OFF állapotú ellenállás egyértelmű korrelációt mutat az írási idővel, az írási idő elhangolódása tehát megmutatkozik az OFF állapotú ellenállások korrelált driftjelenségében is. Ezzel szemben az ON állapotú ellenállás teljesen függetlennek mutatkozik. Ezek az eredmények azt sugallják, hogy mintegy 100, azonos meghajtással előidézett kapcsolási cikluson keresztül viszonylag jól reprodukál a rendszer OFF állapota, és ennek megfelelően egy jól definiált karakterisztikus idővel rendelkező exponenciális eloszlás írja le a sztochasztikus bekapcsolási időket. Nagyobb számú kapcsolás esetén azonban megváltozik (driftel) az OFF állapot mikroszkopikus szerkezete, és ennek megfelelően az írási idők karakterisztikus skálája, ami miatt az exponenciális írási idő eloszlás lognormális eloszlássá szélesedik.

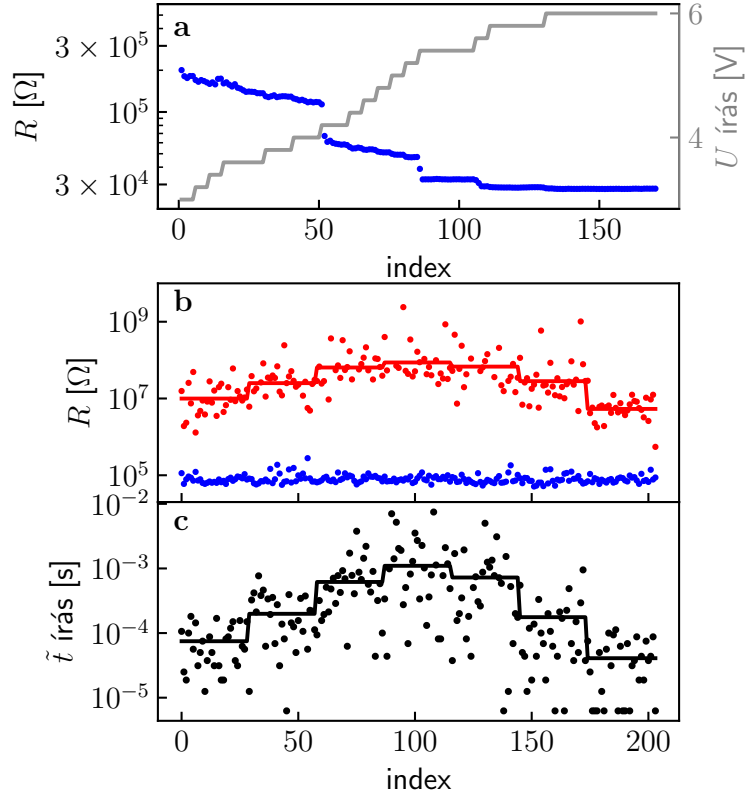
2.3.2. Többszintű programozhatóság

Már a korábbi vizsgálatokból ismert volt, hogy az SiO_x alapú memrisztorok ON állapotú ellenállása széles tartományban változhat az írásra használt feszültségimpulzus amplitúdójának és hosszának függvényében. Jómagam TDK munkámban azt vizsgáltam, hogy az íróimpulzusok segítségével folytonosan változtatható-e az eszköz ON állapotú ellenállása. Ezt a finomhangolási lehetőséget *többszintű programozhatóságnak* nevezzük. Feltérképezéséhez egy $R_{\text{ON}} \approx 200 \text{ k}\Omega$ -ban előkészített mintára fokozatosan növekvő amplitúdójú író és egységnyi amplitúdójú olvasó impulzuspárokat adtam, többször egymás után. A gerjesztő jel és a mért adatsor a 2.6.a ábrán láthatóak.

Megállapítható, hogy egy nagyságrendnyi tartományban lehetett finomhangolni az eszköz ellenállását. Természetesen az is látható, hogy a mért ellenállásértékekben a célzott fokozatosság ellenére is megfigyelhetők véletlen, ugrásszerű változások. Az analóg hangolhatóságot igénylő eszközökben ezek a nemideális hatások gyakran kiküszöbölhetők körültekintő tervezéssel, vagyis a grafén elektródás SiO_x memrisztor alkalmas lehet folytonos hangolhatóságot igénylő hardveres feladatokra.

2.3.3. Kereszteffektusok

A korábbi tapasztalatok alapján a kapcsolási idők eloszlása az írófeszültség változtatásával sok nagyságrendet átfogó tartományban hangolható. Az előzőekben pedig láttuk,



2.6. ábra. Többszintű programozhatóság és kereszteffektusok. **a)** Az ON állapot határán preparált minta lépcsőzetes ellenállásváltozása (●), fokozatosan növekvő amplitúdójú íróimpulzus (—) hatására. **b)** ON állapotú ellenállás (●) és ciklusonkénti logartimikus átlaga (—), illetve OFF állapotú ellenállás (●) és ciklusonkénti logaritmikus átlaga (—). Az ON állapotú ellenállások jó közelítéssel állandóak, míg az OFF állapotú ellenállások egyértelmű korrelációt mutatnak a törlőfeszültség amplitúdójával. **c)** Írási idők (●) és logaritmikus átlaguk (—). A törlő impulzussal való korreláció itt is szépen megfigyelhető.

hogy az írási idő az OFF állapotú ellenállással állandó gerjesztőparaméterek mellett korrelált. Felmerül a kérdés, hogy a mérőjel további beállításai szintén hatnak-e az eloszlásra. TDK munkámban megvizsgáltam, hogy a törlő feszültség, az OFF állapotú ellenállás, és az írási idő-eloszlások között kimutathatóak-e esetleges összefüggések. Ehhez feszültség-impulzusokat kapcsoltam a mintára, majd minden harmincadik mérés után megváltoztattam a törlőfeszültség-amplitúdót. Ezzel fél voltos lépésekben mindkét irányban bejárva a 8.5 V és 10 V tartományt a 2.6.b és a 2.6.c ábrákon látható adatsorokat rögzítettem.

Itt látható, hogy a törlőfeszültség amplitúdójával egyértelműen korrelált az írási idő. A törlőfeszültség növelése tehát késlelteti a következő lépésben a kristályos mag kialakulásának várható időpontját. Megfigyelhető továbbá, hogy a folyamatban az OFF állapotú ellenállások szintén hangolhatóak a törlő feszültséggel, míg az ON állapotú ellenállások átlagértéke jó közelítéssel állandó. Azt tapasztaltam tehát, hogy a törlő impulzussal nemcsak a kapcsolási idők várható értéke, hanem a várható OFF állapotú ellenállások is összefüggést mutatnak. Ezen túlmenően a hangolhatóság reprodukálható volt növekvő és csökkenő amplitúdók esetén is, így kizárható egy egyszerű időbeli eltolódás jelenléte.

Így tehát az OFF állapotú ellenállás és az írási idő között spontán (2.5.d) és szabályozott (2.6.b) korrelációt is sikerrel mutathattam ki.

A grafén elektródás SiO_x memrisztor tehát jól meghatározott eloszlást követő kapcsolási idejének és a fellelhető keresztteffektusoknak köszönhetően alkalmas véletlen szám generálási feladatok ellátására. Ezentúl analóg hangolhatóságot is mutat, ami predesztinálja crossbar struktúrába való integrálásra. Ezen koncepcióról és az eszköz lehetséges integrációjának szimulációjáról a következő fejezetben írok részletesen.

3. Komplex problémák megoldása memrisztorhálózatokkal

Az integrált áramkörök gyártásában az atomi méretskálák elérésével a Moore-törvény végéhez közeledünk: a számítástechnikai hardverek gyártásában már nem lehetséges további méretcsökkentés. Alternatív megoldásokra van szükség, melyek koncepcionális újdonságokkal szolgálnak. A memrisztorok épp ilyen új megoldásokat kínálnak, különösképp a memóriában történő műveletvégzés és a biológiai alapú számítások hardveres kivitelezése terén [10]. Ezek fontosságát mi sem példázza jobban, minthogy alkalmasak a műszaki életben oly gyakran előforduló NP-nehéz minősítésű feladatok megoldására. Ebben a fejezetben, a memrisztoralapú hardverek megvalósításának elméletét, a bennük rejlő lehetőségeket és az ezekkel elért eredményeket ismertetem, kiegészítve az általam készített szimulációk bemutatásával.

3.1. Neurális hálózatok

A legígéretesebb tulajdonsága egy új, memrisztoralapú adattárolási paradigmának, hogy kiküszöbölheti a von Neumann architektúra legnagyobb hiányosságát. Klasszikusan ugyanis az adatokat mozgatni kell a központi feldolgozóegység és a memória között. Mivel ez a legenergiaigényesebb hardveres műveletek egyike, jelentős energiahatékonyságbeli növekedést eredményezne az adattárolás helyén történő műveletvégzés. Mivel a rezisztív kapcsoló memóriák vezetőképességükben kódolják az információt, a tárolt érték fizikailag egy helyen olvasható vagy modulált olvasó feszültségimpulzussal skálázható, tehát az egy helyben történő tárolás és számítás lehetősége biztosított [9]. Sőt, a memrisztorokat *crossbar* struktúrának nevezett mátrix sémába rendezhetjük. Velük így egy lépésben végezhetjük el egy N elemű vektor $N \times M$ nagyságú mátrixszal való összeszorozását, ami klasszikusan $N \times M$ szorzás és akkumulálás műveletet igényelne. Ezentúl pedig, mivel a neurális hálózatok működése matematikailag mátrix-vektor szorzások összességére vezethető vissza, a crossbar elrendezés ideális jelölt a neurális hálók hardveres megvalósítására is.

3.1.1. Crossbar struktúra

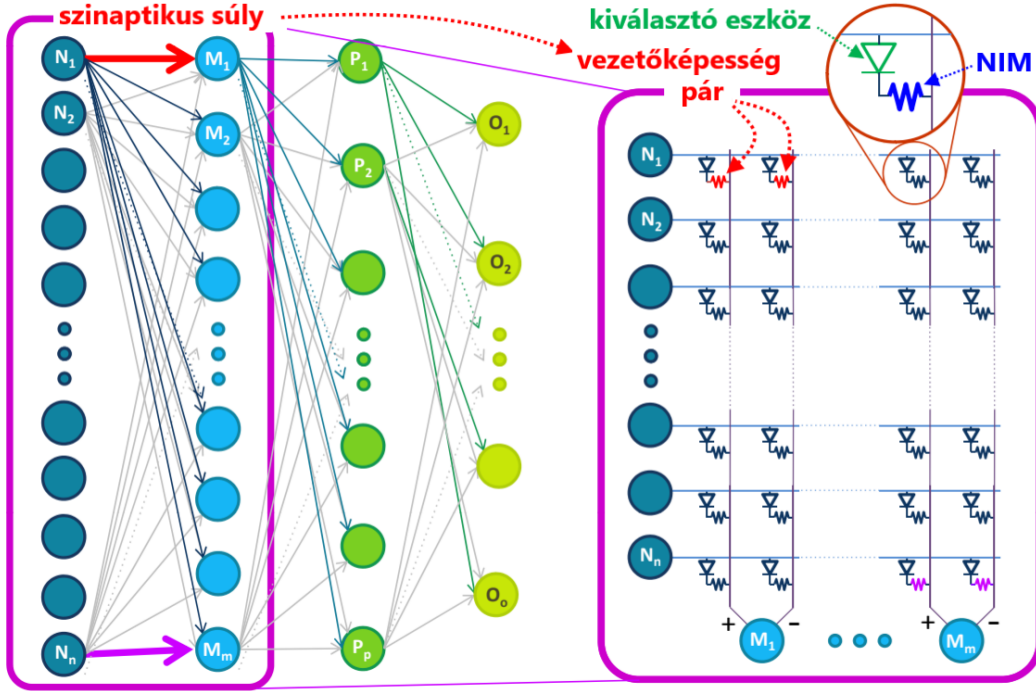
A crossbar elrendezés egy fizikai mátrix séma, melynek egyes mátrixelemei memrisztorból vagy memrisztorokból álló egységek. Egy ilyen mátrix modellje látható a 3.1. ábra jobb oldalán. Azért, hogy az egyes mátrixelemekhez célzottan férhessünk hozzá és hogy a vezetékekben ne folyhasson parazita áram, ezek gyakran kiválasztó tranzisztorokon segítségével címezhetőek.

Amikor egy lineáris algebrai értelemben vett $\underline{W}$ mátrixot szeretnék átírni a hardveres mátrix sémába, formálisan definiálnunk kell egy $g()$ függvényt. Ez határozza meg, hogy melyik valós számhoz milyen kísérleti vezetőképesség-értéket rendelünk. Bináris (0-1 értékeket tartalmazó) $\underline{W}$ mátrix esetén praktikus szokás az alacsony és magas vezetőképesség-sávokat, vagyis az ON és OFF állapotok vezetőképesség-értékeit megfeleltetni a számoknak. A konverziós függvény meghatározását és a memrisztorok írását követően az eszközökből kirajzolódik egy $\underline{G}$ fizikai vezetőképesség-mátrix. Ebben az egyes

mátrixelemeket a memrisztor(ok)ból álló memóriacellák jelentik, melyek vezetőképessége

$$G_{ij} = g(w_{ij})$$

összefüggés szerint meghatározott. A hardverben az i . sor j . oszlopában található memóriacella eredő vezetőképessége a $\underline{W}$ mátrix i . sorának j . elemét reprezentálja. A tipikus elrendezésben a mátrix egységcellái vízszintes és függőleges szomszédjaikkal kiválasztó egységeken keresztül csatoltak, így szabályozható, hogy a huzalozáson keresztül épp melyik mátrixelem(ek)hez férünk hozzá.



3.1. ábra. Neurális hálózat és sematikus memrisztormátrix megfeleltetése [11]. Egy több rejtett réteget tartalmazó hálózatban kettő réteg neuronjait összekötő szinaptikus súlyok megfeleltethetők egy mátrix elemeinek. A neuronokon áthaladó értékeket feszültségértékekkel reprezentálva egy súlymátrix leképezhető fizikai memrisztormátrixra, crossbar sémába rendezett rezisztív kapcsoló memóriákra. Az ábrán szereplő hálózat esetében egy súlytényezőt kettő memrisztor reprezentál, mindkettő nem-illékony memóriaegység külön kiválasztóelektronikán keresztül férhető hozzá.

A fizikai mátrix vízszintes huzalain keresztül feszültséget kapcsolhatunk a mátrix egyes soraira. Matematikai megközelítésben a vízszintes huzalokon eső feszültségek egy n elemű $\underline{N}^T$ bemeneti oszlopvektort alkotnak, amit bemeneti vektornak nevezünk. A bemeneti vektort az elrendezésre kapcsolva N_i feszültség esik a mátrix i . sorának párhuzamosan kapcsolt elemein. Rajtuk ennek hatására függőleges irányban az Ohm-törvény szerint $N_i \cdot G_{ij}$ áram kezd folyni. Ha több bemenetet is aktiválunk, j . oszlopban található memrisztorokra ez a függőleges csatolás révén a Kirchhoff-törvény értelmében összegződik. A függőleges huzalokon eső feszültségek pedig egy m elemű $\underline{M}$ kimeneti sorvektort határoznak meg. A j oszlopban található mátrixelemeken átfolyó áramok az M_j vektorelemében

akkumulálódnak, benne áram mérhető. Ennek értéke a leírtak értelmében

$$M_j = \sum_{i=1}^n N_i \cdot G_{ij},$$

ami a mátrix-vektor szorzás jól ismert összefüggése. Analóg módon programozható memrisztorokat crossbar struktúrába rendezve tehát fizikailag is képesek vagyunk tetszőleges mátrixot reprezentáló hardvert építeni, amely a szorzások végeredményét egyetlen lépésben, elemi fizikai törvényekre alapozva határozza meg.

3.1.2. Hardveres implementálás

Mind az építőelemek, mind pedig a crossbar elrendezés minden előnyös tulajdonságát kiaknázzák azok a törekvések, melyek neurális hálózatokat igyekeznek hardveresen megvalósítani. A neurális hálózatok egy alcsoportját képező *mély neurális hálózatot* láthatunk a 3.1. ábra bal oldalán. Itt megfigyelhetjük, hogy ezek bemeneti, kimeneti és egy vagy több köztes neuroncsoportból állnak. A köztes csoportokat rejtett rétegnek nevezzük, az egyes neuronokat pedig szinaptikus súlyok kötik össze, amik a hálózatban terjedő adategységeket szorozzák. A hálózat matematikai modelljében a neuronok egyes csoportjaihoz egy-egy vektort, a rétegeket összekötő szinaptikus súlyokhoz pedig egy-egy mátrixot rendelhetünk. A hálózat működése ezáltal sorozatos mátrix-vektor szorzásokra egyszerűsíthető, működtetése pedig egyértelműen végezhető memrisztoralapú crossbarokat tartalmazó hardveres egységekben.

A mély neurális hálózatok egyik fő alkalmazási területe a *klasszifikációs feladatok* megoldása. Itt egy bemenő adathalmaz elemeit kell előre meghatározott csoportokba beosztani. Egy elem besorolásakor az elemet reprezentáló vektor a hálózat bemenetére kerül. Kézzel írott számjegyek csoportosítása során ezt úgy képzelhetjük el, hogy ha a számjegy egy 22×24 pixeles képen látható, ahol minden pixel a hálózat egy-egy bemeneti neuronjának feleltethető meg. Ezekből az adat szálakon keresztül jut a következő réteg neuronjaihoz. A szálak szinaptikus súlyaikkal felszorozzák, a fogadó neuronok pedig akumulálják a beérkező értékeket, majd konstans értékkel eltolják illetve megadott függvény szerint transzformálják őket. Az így kapott értékek a rejtett rétegeken keresztül jutnak el a hálózat végét jelentő záró rétegig. Innen olvasható ki az Q kimenet, a klasszifikáció végeredménye, ami a példánkban a $0, 1, \dots, 9$ számjegyek egyike lenne.

A besorolás alapját a szinaptikus súlyok hangolása, a neurális hálózat betanítása adja. Ekkor a kezdeti súlytényezőket egy tanító adathalmaz és egy meghatározott szabály szerint változtatjuk, fokozatosan növelve a klasszifikációs pontosságot. Ahhoz, hogy ezt is hardveresen végezhessük, analóg hangolhatóságot mutató memrisztorokból álló mátrixra van szükség. Az elrendezés kiválasztó egységeinek épp ekkor, a vezetőképesség-mátrix programozásakor, vagyis a neurális hálózat a hálózat tanításakor van szerepe. A sikeresen tanított neurális hálózat működésekor, már az összes memrisztoregység használatban van, így a kiválasztó eszközöknek ebben a működési fázisban nincs további feladata.

Felmerül természetesen a kérdés, hogy a fizikai mátrixelemek vezetőképességének szórása, illetve az analóg hangolhatóság hogyan befolyásolja az idealizált hálózatok működését. A dedikált mérésekből az derült ki, hogy a memrisztor alapú neurális hálózatok eredendően is magasan hibátűrőek. Több tanulmányban is kimutatták, hogy ha a hálózatot alkotó eszközök akár 20%-a véglegesen alacsony vagy magas vezetőképességű állapotban marad, a súlyok ismételt betanításával majdnem teljesen visszaállítható az eredeti

klasszifikációs pontosság [12]. A fennmaradó limitáló tulajdonságok pedig körültekintő eszköztervezéssel ellensúlyozhatónak mutatkoznak. 2018-ban például az egyirányú és néhol nehezen kontrollálható súlyfrissítés kiküszöbölésére prezentáltak egy speciális szinaptikus egységcella elrendezést. Kettő, eredő hatásában ellentétesen szereplő fázisváltó egységet kombináltak CMOS eszközökkel [11], a memrisztorokból felépülő mátrixot pedig kísérletileg is megvalósították. A CMOS eszközök nemidealitását azok kimért és validált karakterisztikája alapján szimulálták. Kézzel írott számjegyek felismerésében az MNIST betanító és tesztelő adatbázis alapján végzett betanítás után a tesztképek klasszifikációjában a szimulált hálózat 97.95%-os, míg a magasan optimalizált szoftveres TensorFlow csomag 97.94%-os klasszifikációs pontosságot ért el.

3.2. Valószínűségi optimalizálás Hopfield-hálózatokkal

A neurális hálózatok egy potenciális felhasználási területe az NP-nehéz feladatok heurisztikus megoldása. A műszaki gyakorlatban előforduló problémák esetében a cél egy rendszer leíró paraméterein definiált energiafüggvény minimalizálása. Egy integrált áramkör esetén a feladat lehet például a fő funkcionális blokkok elhelyezése egy felületen, úgy, hogy az őket összekötő vezetékek teljes hossza (az energiafüggvény) minimális legyen. Milliós nagyságrendű elhelyezendő alkatrészs szám esetén ez a probléma meglehetősen bonyolulttá válik.

Persze kézenfekvő megoldási lehetőségnek tűnik a paramétertér fokozatos, de véletlenszerű sorrendben történő végigléptetése, ahol egy lépést - a kiválasztott paraméter értékének megváltoztatását - akkor végzünk el, ha ez az energiafüggvényt csökkenti. Ilyenkor szemléletesen a függvény által meghatározott energiafelületen próbálunk mozogni, keresve a legmélyebben fekvő pontot. Azonban az, hogy az energiát csak csökkenteni tudjuk, vagyis hogy csak lefelé lépdélhetünk a felületen, gyakran okoz lokális minimumhelyen való ragadást. A cél természetesen a globális szélsőérték hely azonosítása, ezért dolgozták ki a *szimulált hűtési* (*simulated annealing*) eljárást, melynek alapötlete, hogy véges valószínűséggel elvégezzünk olyan lépéseket is, melyek növelik a rendszer energiáját [13]. Ezzel megelőzhetjük, hogy lokális minimumban ragadjon a rendszer, az elfogadási valószínűség folyamatos csökkentésével pedig a megoldás konvergenciája is biztosítható. A szimulált lehűtés a kombinatorikus optimumkeresési feladatok megoldásának megbízható stratégiájaként terjedt el. Azonban tudjuk, hogy a számítógépesen szimulált megoldási módszerekhez képest jelentős előrelépést jelent, energiahatékonyságban és a megoldáshoz szükséges időben is, ha egy feladatot hardverszinten tudunk kezelni. A cél tehát a szimulált lehűtés alapötletének átültetése egy fizikailag megvalósított neurális hálózatba. Ahhoz, hogy ennek a folyamatát feltárhassuk, nézzük meg először a probléma matematikai megfogalmazását.

Egy *kvadrátikus bináris optimalizációs probléma* (*Quadratic Unconstrained Binary Optimization*) azt jelenti, hogy egy rendszer N darab x_i paraméterrel jellemezhető és hozzá

$$f : \mathbb{B}^N \rightarrow \mathbb{R} \quad f_Q(\underline{x}) = \sum_{i=1}^N \sum_{j=1}^i q_{ij} \cdot x_i x_j$$

függvény rendelhető, ahol $\mathbb{B} = \{-1, 1\}$ és $\forall i : x_i \in \mathbb{B}, \forall i, j \ q_{ij} \in \mathbb{R}$ pedig a probléma által meghatározott együtthatók. Az $f()$ függvényt ekkor a probléma energiafüggvényének

nevezzük és célunk az

$$\underline{x} = \underset{\underline{x} \in \mathbb{B}^N}{\operatorname{argmin}} f_Q(\underline{x})$$

egyenlet megoldása. Az energiafüggvény a $\underline{Q} \in \mathbb{R}^{N \times N}$ szimmetrikus együttthatómátrix bevezetésével $f_Q(\underline{x}) = \underline{x}^T \underline{Q} \underline{x}$ alakban is írható.

A Hopfield-hálózat egy neurális hálózat ami egy *irányítatlan és súlyozatlan egyszerű gráffal* írható le. A gráf jellemezhető $\underline{W}$ *szomszédsági mátrix*ával, ami súlyozatlan mátrixról lévén szó, bináris lesz. Sőt, ezen reprezentációban a mátrix definíció szerint szimmetrikus $w_{ij} = w_{ji}$ és a diagonálisában álló elemek nullák $w_{ii} \equiv 0$. Ez azt jelenti, hogy a súlytényezők értéke haladási iránytól független, hiszen a gráf irányítatlan, és egyszerű gráfról lévén szó egyetlen neuron sem csatolt önmagával. A bemeneti neuronok állapotát $\underline{x} = [x_1, \dots, x_N]$, $x_i \in \{-1, 1\}$ vektor írja le.

A hálózatot kezdetben *asszociatív memóriaként* használták, vagyis feladata keresett adat annak darabjaiból való előállítása volt. Működtetése a *frissítési szabálya* szerint történt, ami azt mondja ki, hogy az t . lépésben a k . neuron x_k állapota

$$A_k = \sum_{j \neq k} w_{jk} x_j \quad x_k^* = \begin{cases} +1 & \text{ha } A_k \geq \theta_k, \\ -1 & \text{ha } A_k < \theta_k, \end{cases}$$

szabály szerint módosul, ahol θ_k pedig az k . neuronhoz tartozó küszöbérték. A rendszerről belátható, hogy a frissítésekkel az

$$f(\underline{x}) = -\frac{1}{2} \sum_{i,j=1}^N w_{ij} \cdot x_i x_j + \sum_{i=1}^N \theta_i x_i$$

energiafüggvényt csökkenti, ennek bizonyítása az A. függelékben olvasható. A

$$w_{ij} = -q_{ij} \quad \theta_k \equiv 0$$

megfeleltetést alkalmazva tehát könnyen látható, hogy a Hopfield-hálózat alkalmas a kívánt optimalizációs feladat megoldására.

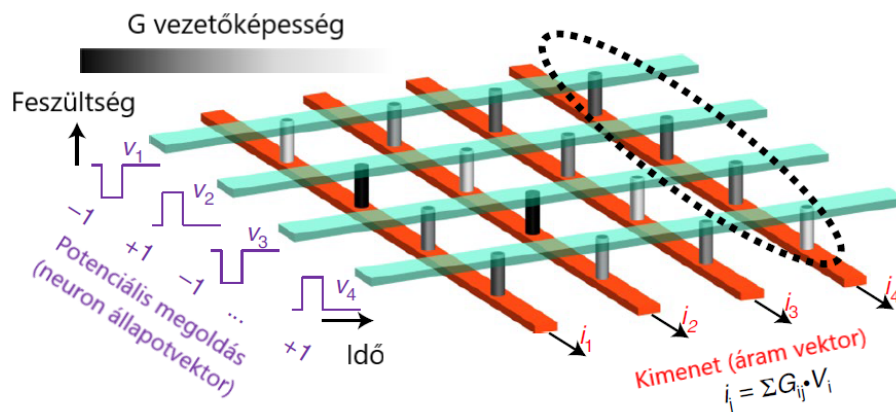
A módszer továbbá kiterjeszthető, hogy vele a szimulált hűtés alapelvét megvalósíthassuk. Itt az energianövelés lehetőségét egy extra paraméter (η_k) figyelembevétele teszi lehetővé, amivel a frissítési szabály

$$A_k = -\sum_{i \neq k} q_{ik} x_i \quad x_k^* = \begin{cases} +1 & \text{ha } A_k + \eta_k \geq 0 \\ -1 & \text{ha } A_k + \eta_k < 0 \end{cases}$$

alakúra módosul, melyben megjelent egy η_k zaj kontribúció. Itt az új paraméter időbeli fluktuációja tudja pillanatnyilag eltolni a küszöbértékeket, ami biztosítja a sztochasztikusan előforduló energianövekedést. A fluktuáció hangolása, vagyis maximális értékének folytonos időbeli csökkentése az optimális konvergenciához természetesen itt is elengedhetetlen.

3.3. Kísérleti megvalósítás

A hálózat kísérleti megvalósítása 2020-ban sikeresen meg is történt, oxid-alapú, két-terminálos memrisztorokból Hopfield-hálózatként működő fizikai memrisztormátrixot építettek. A megvalósítás részleteit illetve a kísérlet eredményeit [4] forrás alapján mutatom



3.2. ábra. A kísérleti memrisztor Hopfield-hálózat mátrixának sematikus rajza [4]. Fekete-fehér színskála jelzi az analóg hangolhatóságot mutató memrisztor egységelemeket. Lila színnel vannak kiemelve a $[-1, 1]$ elemekből álló bináris vektor elemeit reprezentáló feszültségjelek, míg piros szín jelöli a kimeneti vektort. Ennek egyes elemei az Ohm- és Kirchhoff törvények értelmében összegződő áramértékek az oszlopokra, ami lineáris algebrai értelemben épp a bemeneti vektor és a mátrix szorzata.

be. A hálózat sematikus rajza a 3.2. ábrán látható, ahol megfigyelhetjük a szinaptikus súlyokat vezetőképességükben kódoló memrisztorokat, illetve a be- és kimeneti vektorokat reprezentáló feszültségimpulzusokat. A neuronok frissítését a periférikus elektronika valósítja meg: egy vezérlő egység a hálózatra kapcsolja a bemeneti feszültségeket, a kiolvasni kívánt k . áramérték egy multiplexerrel választható ki. Ebben a hibrid, analóg-digitális memrisztor hálózatban a zajjárulék előállításáért és hangolásáért egy dedikált elektronikai egység és egy számítógép felel. Az egység a memrisztormátrix elektronikai zaját és a vezetőképesség hangolási pontatlanságát kihasználva nyert sztochasztikus jeleket skálázza, majd számítógépesen történik az aktivációs függvény számítása. A neuronfrissítés utolsó lépéseként pedig egy újabb vezérlőelem a bemenetre kapcsolja a kiszámított $\underline{x}^*$ vektort, megkezdve ezzel a következő iterációs lépést.

3.3.1. A sztochaszticitás szerepe

Ahhoz, hogy kiderüljön, hogy a zajjáruléknak van-e szerepe az iterációs folyamatban és hogy az elrendezés érdemes-e a kísérleti megvalósításra, minősítő feladatot végeztek szoftveresen. Egy a zajjal terhelt Hopfield-hálózatot szimuláló kóddal oldottak meg *max-cut* feladatokat. Ennek célja a gráfok maximális vágásának megtalálása, vagyis keressük a gráf olyan particionálását, melyben a kapott két algráfot összekötő élek száma maximális. A probléma matematikai megfogalmazásáról és optimalizációs feladattá való átalakításáról a B. függelékben írok részletesebben. A feladatok forrása egy *benchmark* könyvtár volt, ami 60, 80 és 100 csúcsú gráfokat tartalmaz a releváns feladatok megoldásával együtt.

Azért, hogy lássák, a zajnak egyáltalán van-e szerepe a megoldási módszer optimalizálásában, a legegyszerűbb zajprofillal kezdték a szimulációkat. Ez a konstans zaj, ekkor

$$\eta_{\max} \equiv \eta_0,$$

vagyis a t . iterációs lépésben sorsolt k . neuron léptetése előtt szintén sorsolnunk kell egy $\eta_{k,t}$ zajjárulékot is, amiről tudjuk, hogy

$$\eta_{k,t} \in [-\eta_0, \eta_0].$$

Ezzel a módszerrel 60×60 mátrixokra $\eta_0 \in [0, 1.5, 4.25]$ konstans zajprofilok mellett vizsgálták a hálózat konvergenciáját. Egy-egy futás végeredménye a 3.3.a ábrán látható. Itt a várakozásoknak megfelelően azt látjuk, hogy zaj nélkül a hálózat lokális minimumban ragad az energiafelületen, túl nagy zaj esetén bár eléri az optimális megoldást, de a nagy fluktuáció miatt nem stabil a megoldás. $\eta_0 = 1.5$ esetén azonban sikerül a globális minimumot elérni és megtartani, látjuk tehát, hogy a Hopfield-hálózatnak szüksége van zaj becsatolásra az ideális működéshez és hogy létezik ideális zajszint.

Az optimális zajszint egy adott optimalizációs probléma esetében könnyen meg is határozható. Ehhez 1000 véletlenszerűen sorsolt bináris vektorból álló könyvtárat készítettek és az $\eta_0 \in [0, 5]$ tartományban $\Delta\eta_0 = 0.25$ lépésekben haladva adott konstans maximális zaj mellett megoldották a kiválasztott max-cut problémát minden vektorból, mint kiindulási állapotból indítva a rendszert. Így mérték az átlagos, minimális és maximális elért energiát az egyes zajszintekre, illetve meghatározták a konvergenciavalószínűséget is. Ez N különböző vektorból indított futás esetén, mindegyik állapotból K iterációs lépést végezve definíció szerint

$$\mathbb{P}_{\text{konvergencia}} = \frac{\text{sikeres konvergenciák száma}}{N}.$$

Ennek a tesztnek a végeredménye a 3.3.b ábrán látható. Egyértelműen leolvasható, hogy az átlagos energiának minimuma, a konvergenciavalószínűségnek pedig maximuma van $\eta_0 \approx 1.5$ értéknél, az optimális zajszintet tehát sikeresen megtalálták.

A szimulált hűtés alapelvét követve azonban általánosíthatjuk a lehetséges zajprofilokat. Jobb teljesítményt várunk ugyanis egy időben lecsengő zajkarakterisztikától, ami a későbbi iteráció lépésekben csak csekélyebb fluktuációknak ad teret. Ekkor az eddigiekkel ellentétben

$$\eta_{\max} = \eta_{\max,t}.$$

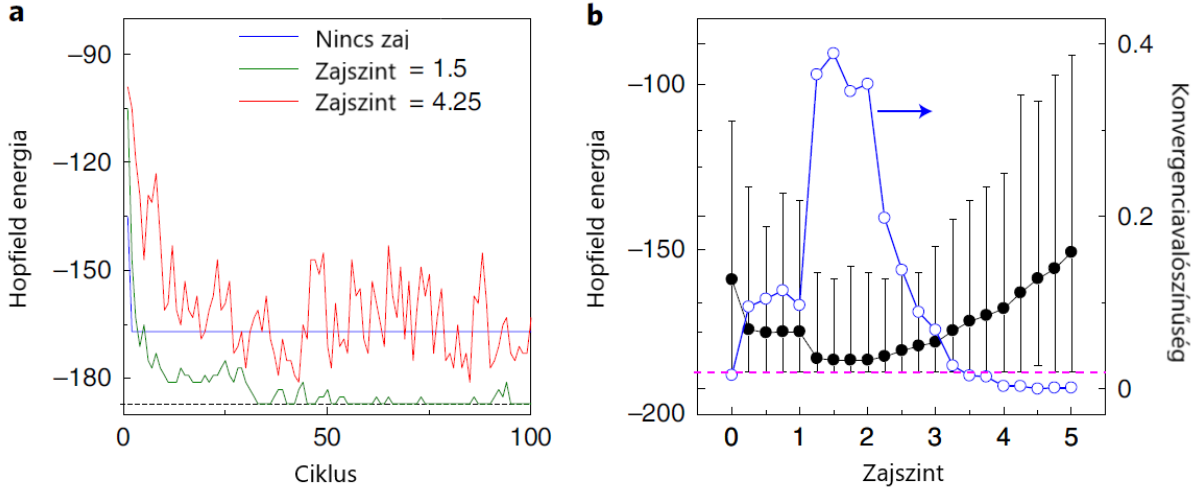
Bár szoftveresen tetszőleges függvényt léptethetünk iterációnként, kísérletileg limitáltak a rendelkezésre álló zajforrások, főleg, hogy az időbeli hangolhatóság igénye is fellép.

3.3.2. A sztochaszticitás forrása

Az elkészített eszközben az η zajjárulék kettő potenciális forrását azonosították. Az első a nanoelektronikai egységekre jellemző *vezetőképesség fluktuáció*, fizikai értelemben vett zaj, ami az atomi fluktuációk és elektronok csapdázódásának eredménye [14]. A második pedig a memrisztorok véges pontosságú hangolhatóságából és a rendszer egyéb nemidealitásából fakadó *számítási hiba*. Ugyanis amikor a

$$w_{ij} \in \{0, 1\} \xrightarrow{g()} \{G_{\text{kis}}, G_{\text{nagy}}\}$$

megfeleltetés alapján elvégezzük a mátrix felprogramozását a kísérletileg megvalósított vezetőképességek szórni fognak várható értékük körül. Vagyis az i . sor j . oszlopában szereplő vezetőképességet csak $\Delta G_{ij} = g(w_{ij}) - G_{ij}$ hiba mellett tudjuk programozni.



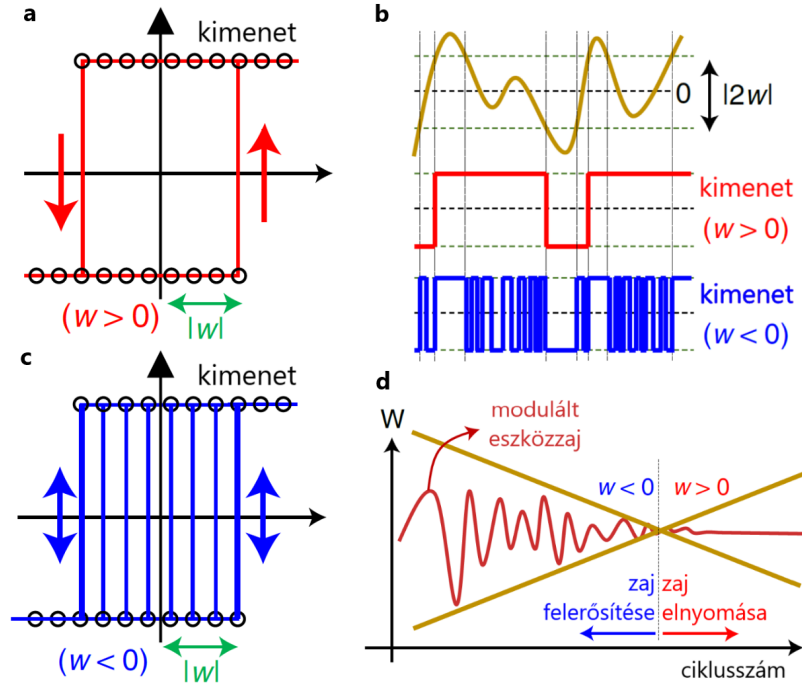
3.3. ábra. Konstans maximális zajjárulék mellett szimulált megoldások [4]. **a)** Három eltérő zajszint mellett látott konvergencia az ismert energiaminimumhoz. Látható, hogy létezik optimális zajszint, ami segít megoldani a feladatot, de nem destabilizálja a hálózatot a globális minimum állapotában. **b)** Az optimális zajszint megkeresése: minden η_0 zajszint mellett megvizsgálták a hálózat konvergenciáját az 1000 véletlen vektorra. Látható a minimális, keresett Hopfield-energia (---), az átlagos energia (●), illetve az elért maximális és minimális energia hibaszáv formájában. Megfigyelhető ezentúl a konvergencia valószínűség (○), ami az átlagos energiához hasonlóan $\eta_0 \approx 1.5$ körül veszi fel extrémumát.

Az előálló vezetőképesség-mátrix tehát alkalmas a magas és alacsony értékek szétválasztására, de a műveletvégzés során véletlenszerű ütemben csatolódik hiba a rendszerbe. Amikor léptetjük a hálózatot, akkor a random módon kiválasztott k . oszlop határozza meg, hogy éppen mely, egyenként adott ΔG_{ij} hibájú súlyokból számoljuk az aktivációs függvényt, sztochasztikussá téve az idealizált és valós mátrixokkal végzett számítások végeredményének eltérést.

A rendszer működése közben mindkét sztochasztikus komponens együttesen fejti ki a hatását, dedikált mérősorozattal azonban sikeresen izolálták a két ismertetett fő komponenszt. A valódi fluktuációkból származó dinamikus zajjárulék méréséhez azonos bemeneti vektorokat adtak ismételtén a hálózatra. Közben mérték a crossbar-oszlopok időben változó áramát, melynek ingadozása μA nagyságrendűnek adódott. Az azonos bemenetekre adott eltérő kimenet természetesen a pillanatnyi fluktuációk eredménye.

A számítási hiba hatásának vizsgálatához a crossbar-struktúrára időben változó bemeneteket adtak, mérve az áramválaszt. Közben meghatározták, hogy egy ideálisan felprogramozott vezetőképesség-mátrix használatával milyen eredményt kaptak volna. A mért és optimális mátrix esetében kapott kimenetek eltéréseiből meghatározható a számítási hibák nagyságrendje, ami mA tartományba esett. Ez bizonyult tehát a rendszerben domináns zajkomponensnek. Bár nem egyszerű várható érték körüli ingadozást látunk, a véletlenszerű léptetés és így sztochasztikus számítási hibák megjósolhatatlanul ingadoznak időben. Ezentúl a crossbar struktúrába rendezett memrisztorok aritmetikai hibáiról kimutatták, hogy erősen nemlineárisak, bemenetfüggőek és Gauss-eloszlású számítási hibát eredményeznek [15].

Mivel azonban maga az iterációs folyamat során változatlan crossbar elrendezés az ere-



3.4. ábra. A *hysteretic threshold* működésének szemléltetése [4]. **a, c)** A kimenet a bemenet függvényében kisimul ($w_t > 0$, —) vagy zajosabbá válik ($w_t < 0$, —) a hiszterézisnek köszönhetően. **b)** Állandó w_t esetén mérhető η_t válaszfüggvények egy példa bemenetre (—). **d)** Példa hangolt w_t paraméterrel elért csökkenő $\eta_{max,t}$ (—) mellett generált η_t zajfüggvényre (—).

dő zaj fő forrása, az ebből nyert zajparaméter η_{max} maximális értéke továbbra is állandó. A *lehűtésért*, vagyis az időben folyamatosan csökkenő zajamplitúdóért az újdonságként prezentált *hysteretic threshold* algoritmus felel, melynek célja biztosítani, hogy a globális minimum elérését követően jelentős elmozdulás ne történhessen a probléma energiafelületén. Így a η_k paraméter már időfüggést mutat, vagyis

$$\eta_{k,t} = \eta_k - w_t(x_{k,t}).$$

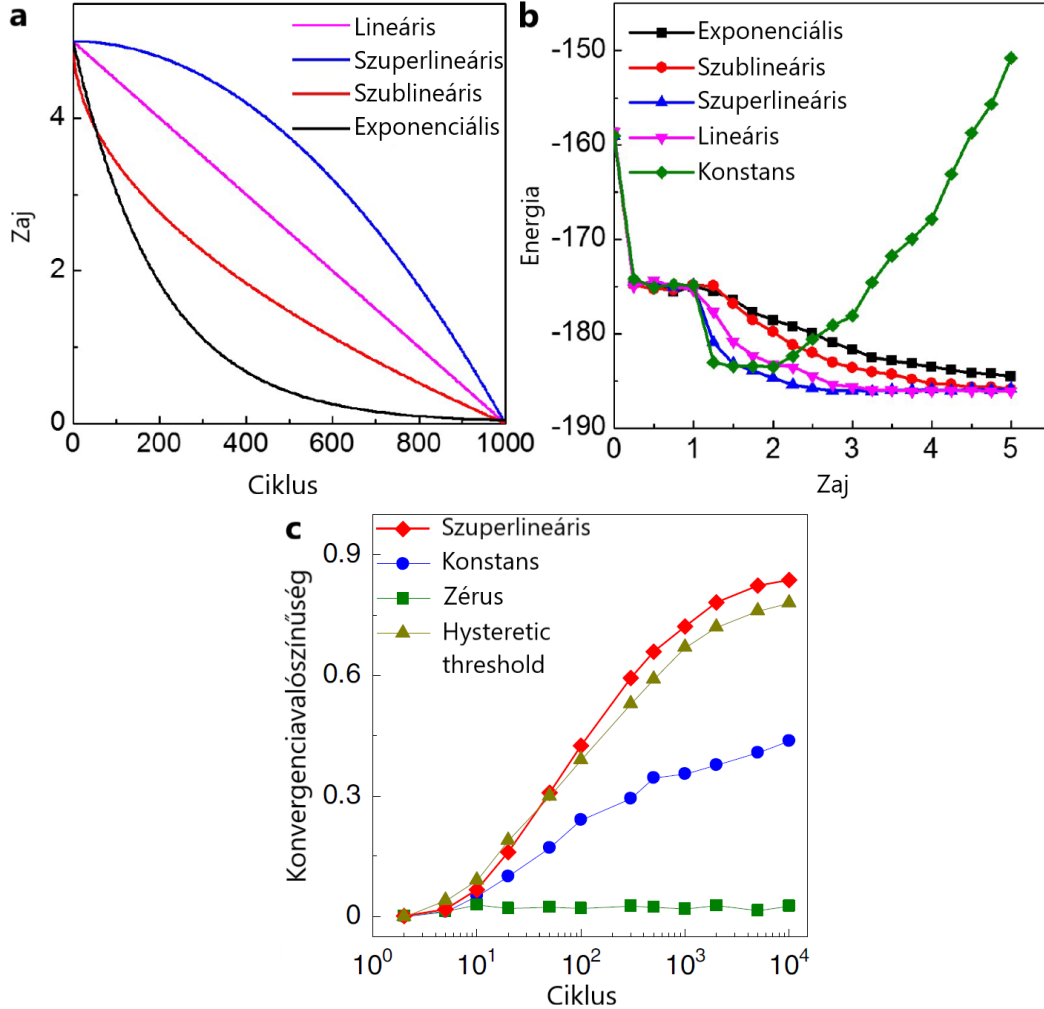
Az új $w_t()$ függvény Schmitt-trigger jellegű működést tesz lehetővé, ami annyit jelent, hogy a fluktuáló $x_{k,t}$ feszültségérték egy nulla körüli $2|w_t|$ széles tartományba esik, akkor w_t előjelének függvényében képes az elektronika a fluktuációt felerősíteni vagy elnyomni. Működésének szemléltetése a 3.4. ábrán látható, de tekintsük át egy példán keresztül is: tegyük fel, hogy az $x_{k,t}$ feszültségérték nagyobb kezdetben, mint w , de folyamatosan csökken. Amikor értéke w_t alá csökken, kettő eset lehetséges:

- $w_t > 0$ esetén a kimeneti jel megőrzi kezdeti értékét, ezzel a zaj csökken, hiszen a nulla körüli tartományban ingadozó jel fluktuáció nem jutnak a kimenetre. Ha $x_{k,t}$ ezek után $-w_t$ alá csökken, a kimenet alacsony értékre vált, ha pedig nőni kezdene, w_t -t átlépve továbbra sem történik változás.
- $w_t < 0$ esetén azonban a $0 \pm |w_t|$ tartományba érve az eszköz elkezd magas és alacsony kimenetei között - astabil billenőkörre emlékeztetően - váltogatni, felerős-

sítve ezzel a fluktuációt, hiszen konstans $-|w_t| < x_{k,t} < |w_t|$ feszültségjel esetén is ingadozó kimenetet látnánk.

Egy elektronikai egységgel tehát képesek voltak a memrisztorok és a crossbar struktúra saját zaját felerősíteni vagy elnyomni és ezt az optimalizációs feladatok megoldására felhasználni.

3.3.3. A hálózat teljesítménye



3.5. ábra. Az egyes zajprofilok és a hysteric threshold összehasonlítása [4]. **a)** Az optimális zajprofil keresésénél felmerülő lehetséges karakterisztikák $\eta_0 = 5$ kezdeti amplitúdóval és $K = 1000$ lépésre meghatározva. **b)** Konvergencia valószínűség a különböző jellegű profilok esetén. Látható, hogy a szuperlineáris profillal érhető el a minimális átlagos Hopfield-energia. **c)** A kísérleti elrendezés teljesítményének összehasonlítása az ideális karakterisztikával, a kettő esetében a konvergencia valószínűsége a ciklusszám függvényében ugyanolyan ütemben növekszik.

Mielőtt rátérnénk a hysteric threshold egység teljesítményének értékelésére, szükségünk van egy referenciára, amivel összehasonlíthatjuk. Azért, hogy a Hopfield-hálózat

ideális zajprofilját, melyre a konvergenciatulajdonságok a legkedvezőbbek, meghatározhassák, többféle függvény használatával is szimulálták a hálózat működését. Ezek a 3.5.a ábrán láthatóak. A konstans zajprofil ideális amplitúdója meghatározásának módszerével ezúttal az ideális zajprofil keresték. A cikk függelékéből kiderül, hogy a szuperlineáris karakterisztika mellett érhető el a legkisebb átlagos Hopfield-energia, erről a 3.5.b ábra megtekintése révén győződhetünk meg, a továbbiakban tehát az ezzel elért eredmények szolgálnak referenciaként.

A 3.5.c ábrán pedig láthatjuk azt, hogy a futási ciklusok függvényében mekkora a konvergencia valószínűség néhány zajprofilra, illetve az elektronika által megvalósított hysteretic threshold algoritmusra. Szembetűnő, hogy a szimulációkban legkiemelkedőbben teljesítő szuperlineáris karakterisztikával szinte teljes megegyező a konvergencia valószínűsége a hysteretic threshold alkalmazása esetén is, a kísérleti megoldó tehát ideálisnak tekinthető. Ezentúl ismételten leolvasható, hogy a zajjárulék nélkül működő hálózat értelmezhetetlenül gyenge megoldási módszer önmagában, ehhez képest már a konstans zaj alkalmazása is jelentős javulást eredményez.

A cikkben prezentált eredmények jelentősége kettős. Egyrészt elméleti és kísérleti megalapozását adják a neurális hálózatok rezisztív kapcsoló memóriákkal való hardveres megvalósíthatóságának, kijelölve az irányt a terület további kutatásához. Másrészt pedig mérhetővé teszik a passzív memrisztormátrixokkal való optimalizáció pontosságát és energiaigényét. Mint láttuk, az ideális, szoftveres megoldókkal a sztochaszticitásért felelős elektronikai egység konvergencia valószínűségben közel azonos eredményekre képes. Ami pedig a működtetés energiaigényét és a technikai integrálhatóság kérdését érinti, a kor népszerű kvantummechanikai elveken működő hardveres gyorsítói a fejlettség jelenlegi fázisában magas hűtésigényüknek köszönhetően egyelőre nem tudják szignifikánsra növelni előnyüket a, például a fentiekben is tárgyalt, klasszikus megoldókkal szemben.

3.4. Hopfield-hálózat szimulációja

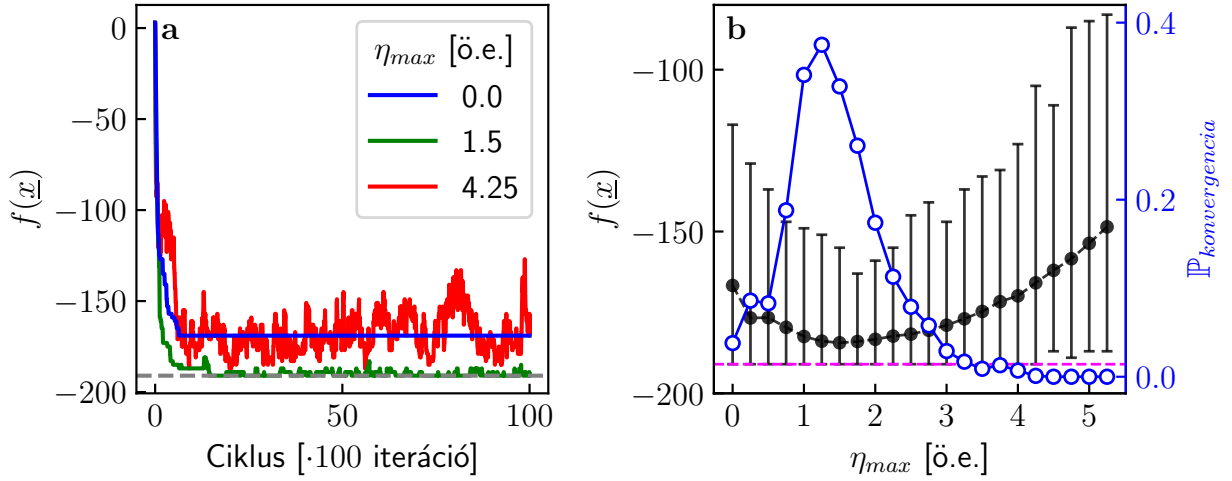
Azért, hogy a szoftveresen működtetett Hopfield-hálózatokat mélyebben megérthessem és reprodukálhassam a bemutatott eredményeket, Python programozási nyelven szimulációs kódot írtam. Ez képes a *benchmark* könyvtár adatait beolvasni és szimulált lehűtés illetve Hopfield-hálózatok segítségével is megoldani a feladatokat. A zaj kontribúció maximumának profilja tetszőleges lehet, jelenleg konstans, szuper- és szublineáris, lineáris és exponenciális $\eta_{\max,t}$ profilok is választhatóak. Ezeket a kód K lépésű iterációra zérus értékhez tartó profilok esetén automatikusan generálja, exponenciális profil esetén pedig a lecsengés τ paramétere változtatható. A futás során a t . lépésben a konkrét $\eta_{k,t}$ paraméter sorsolása szintén több valószínűségi eloszlás szerint történhet. Egyenletes eloszlás esetén a nullára szimmetrikus intervallumnak köszönhetően

$$\mathbb{E}(\eta_{k,t}) = 0 \quad \text{és} \quad \mathbb{D}^2(\eta_{k,t}) = \frac{2 \cdot \eta_{\max,t}}{\sqrt{12}} = \frac{\eta_{\max,t}}{\sqrt{3}}.$$

Ezenkívül elérhető Gauss-féle normál eloszlás is, ezt szintén nulla várható értékkel és az előbb levezetett szórással sorsolom, egymásra skálázva az egyes eloszlásokat.

Saját szimulációs kódom validálására a könyvtár *g05_60.3* jelű gráfját használtam, az itt bemutatott eredmények is az általa kódolt max-cut feladat megoldására vonatkoznak. Elsőként én is készítettem egy 1000 véletlen vektorból álló fájlt, majd megnéztem,

a Gauss-eloszlású, konstans maximális amplitúdójú zaj becsatolása milyen eredményre vezet. A kapott három megoldás a korábbról megismert $\eta_0 \in [0, 1.5, 4.25]$ zajszintekre a 3.6.a ábrán láthatóak.



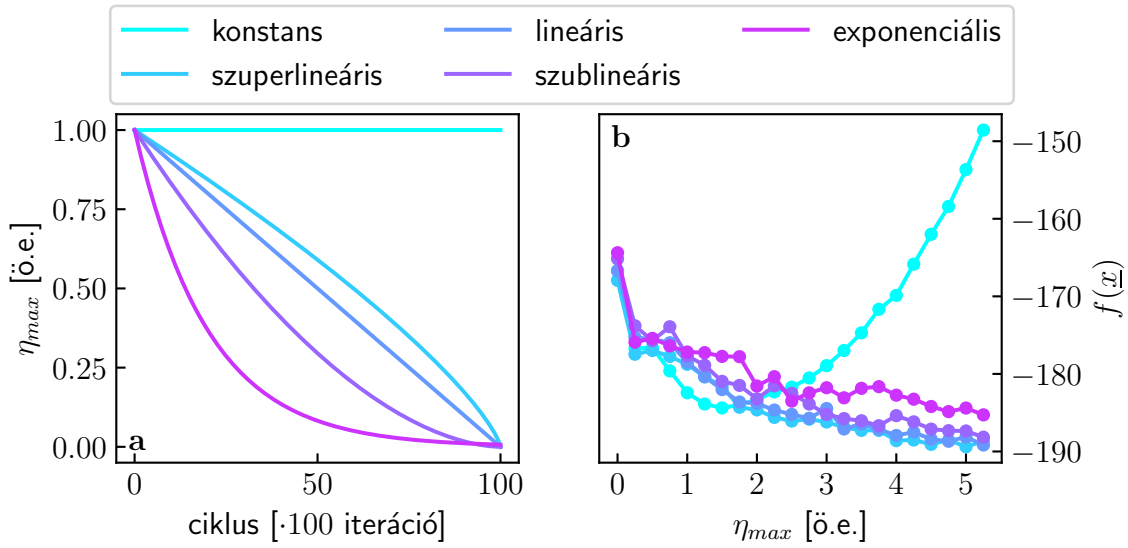
3.6. ábra. Konstans maximális zajjárulék mellett, Gauss-féle normál eloszlással sorsolt zajprofilokkal szimulált megoldások. Az ábra jelölésrendszere megegyezik a 3.3. ábráéval. **a)** A három eltérő zajszint mellett látott viselkedés, az optimális zajszint létezése itt is megfigyelhető. **b)** Az optimális zajamplitúdó értékének megkeresése: minden η_0 érték mellett vizsgáltam a hálózat konvergenciáját az 1000 véletlen vektorra. A tendencia a cikkben olvasottakkal megegyezik, az átlagos Hopfield-energia minimumát, a konvergencia valószínűség pedig maximumát $\eta_0 \approx 1.5$ -nél veszi fel.

Mivel a tapasztaltaim megegyeztek a fő forrásban látottakkal [4], továbbléptem az optimális zajszint megkeresésére. Ehhez én is végigpásztáztam a kóddal az $\eta_0 \in [0, 5]$ tartományt $\Delta\eta_0 = 0.25$ -ös lépésekben, minden zajszinten minden tesztvektorból elindítva az iterációt. Ennek végeredménye a 3.6.b ábrán tekinthető meg. Mivel a cikkben olvasottakkal teljes egyezést tapasztaltam, a kódot validáltnak tekintettem. Sikeresen tudtam tehát a látott algoritmusokat implementáló kódot fejleszteni.

Ennek további bizonyítéka, hogy a forrás függelékében szereplő tesztelési folyamat végeredményét, melyben különböző zajprofilokra vizsgálták, hogy az amplitúdó felskálázása hogyan befolyásolja a működést, szintén elő tudtam állítani a saját kódjaimmal. A teszt elvégzéséhez a használt zajprofilok a 3.7.a ábrán, a végeredménye pedig a 3.7.b ábrán figyelhető meg. Ismét, a forrásban mutatott ábrával és konklúzióval maximálisan megegyező eredményre jutottam: a superlineáris zajprofil eredményezi - megfelelő amplitúdóválasztás mellett - a legjobb konvergenciát.

Ez munkám szempontjából kiemelkedő jelentőséggel bír. A szimulációs kód célja hosszútávon ugyanis annak a leellenőrzése, hogy SiO_x alapú memrisztorok használhatóak-e hardveres zajforrásként. Memrisztor zajforrás esetén az alapötlet az, hogy a ΔI áramzajt a hálózat léptetése során kiszámolt mátrix-vektor szorzás áram dimenziójú végeredményéhez additívan csatoljuk. Itt tehát a zaj amplitúdó az abszolút áramzajjal arányos

$$\Delta I \propto \eta_{k,t}.$$



3.7. ábra. A szimulált zajprofilok és az ezekkel elért átlagos Hopfield-energia. **a)** A szimulált zajprofilok, egységnyi amplitúdó mellett, 10^4 iterációs lépéshez generálva. **b)** A konstans zaj korábban látott viselkedésével való összehasonlításhoz az első 100 vektorra végeztem konvergenciavizsgálatot az új zajprofilokkal. Az átlagos energia tekintetében a szuperlineáris karakterisztika teljesít a legjobban, hajszálnyival megelőzve a lineáris profilt.

Az SiO_x alapú rezisztív kapcsoló memóriák relatív áramzaja

$$\frac{\Delta I}{I} = \beta \cdot \left(\frac{R}{R_0} \right)^\gamma, \quad R_0 \equiv 1 \, \Omega$$

függvénnyel írható le. Itt β és γ tapasztalati együtthatók, R_0 pedig egy fizikai tartalommal nem bíró tényező az együtthatók dimenziótlanítására. Látható tehát, hogy az eszköz ellenállását analóg hangolás útján folytonosan csökkentve, csökkenő amplitúdójú áramzaj nyerhető ki belőle.

Balázs Péter TDK munkájából kiderül, hogy ON állapotú SiO_x memrisztorokra $\gamma \approx 1.72$ [16]. Kihasználva, hogy az R ellenállású eszközön U kiolvasó feszültség esetén átfolyó áram az Ohm-törvény értelmében $I = \frac{U}{R}$, az abszolút áramzajra

$$\Delta I = \beta^* U \cdot R^{\gamma-1}.$$

Ez jellegre, iterációnként csökkentve az ellenállást, épp a 3.7.a ábrán látható, bizonyítottan ideális konvergenciatulajdonságokkal rendelkező szuperlineáris zajprofil eredményezi, amennyiben a minta analóg hangolhatóságát is kihasználjuk. Az ON állapot határán preparálva tehát egy SiO_x mintát és fokozatosan csökkentve ellenállását, a memrisztor várható zajprofilja éppen ideálisnak mutatkozik extra zaj becsatolás mellett működtetett Hopfield-hálózatok üzemeltetésére.

4. Összefoglalás és kitekintés

Szakterdolgozatomban a grafén elektródás SiO_x memrisztorok vizsgálatával foglalkoztam. Ez egy fázisváltó memória eszköz, melyben a szilícium-oxid rétegben létesített aktív tartomány amorf vagy kristályos állapota grafén elektródákon keresztül becsatolt feszültségimpulzusok segítségével változtatható. Az eszköz vizsgálatát, korábbi hazai és nemzetközi eredményekre támaszkodva, másfél éve volt lehetőségem elkezdni. Ezalatt mind az alaptulajdonságokat feltáró kísérleti, mind pedig az alkalmazhatóságot vizsgáló szimulációs munkákat végezhettem. Az itt prezentált eredmények is a kettő eltérő megközelítést igénylő területről valóak, ezeket tehát jelen dolgozatban külön kezeltem.

Először összefoglaltam a fázisváltó memóriákkal kapcsolatos általános ismereteket, majd bemutattam az Atomi és Molekuláris Elektronika Kutatócsoport releváns eredményeit. Ezután az amorf tartomány kristályosodását mint nukleációs folyamatot leíró elméleti modellt alátámasztó kísérleti eredményeimet prezentáltam. Párhuzamosan összegeztem TDK mérésorozatom, illetve a TDK munkám óta végzett kiegészítő mérések idevágó eredményeit, igyekezve részletes képet mutatni egy komplex fizikai rendszerről és motiválva szerepét a jelenleg is zajló számítástechnikai paradigmaváltásban.

Ezután bemutattam, hogyan lehet memrisztorhálózatokat építeni kísérletileg, tárgyalva szerepüket az adattudományban. Végül rátértem a Hopfield-neurális hálózatra, mint az optimalizációs feladatok egy lehetséges metaheurisztikus megoldójára. Részletesen elemeztem egy friss forrást [4], melyben kísérleti és szimulációs úton erősítették meg a hálózat kiemelkedően előnyös tulajdonságait energiahatékonyság és technikai megvalósíthatóság terén. Végül a cikkben bemutatott eredmények reprodukciójával validáltam saját fejlesztésű szimulációs kódom helyességét és rövid kitekintésben vizsgáltam a SiO_x memrisztorok áramzajának karakterisztikáját. Megmutattam, hogy az általam szimulált Hopfield-hálózattal megoldott max-cut problémához szükséges zajprofil kinyerhető SiO_x memrisztorok áramzajának a hálózatba való becsatolásával.

Kutatásom folytatásaként szeretném az eszköz áramzaját teljes körűen integrálni a szimulációs programba, megvizsgálva az eszköz zajforrásként, sőt, később akár crossbar struktúrába integrált mátrixelemként való viselkedését. Ezen túlmenően szeretném folytatni a kísérleti munkát is; a SiO_x memrisztorok holtideje alatt zajló relaxációs folyamat behatóbb vizsgálatával ugyanis újabb részletek tárhatóak fel az eszközök működésének mikroszkopikus folyamatairól.

5. Köszönetnyilvánítás

Szeretném kifejezni köszönetemet Török Tímea Nóra témavezetőmnek, aki felügyelte a munkámat és elvégezte a minősítő mérést a járványhelyzetben. Továbbá segített az ábrák elrendezésének kidolgozásában és elősegítette jelen dolgozat megírását. Külön köszönet illeti Dr. Halbritter András tanár urat tevékenységem általános koordinálásért, az eredmények értelmezésében való közreműködéséért és a kézirat ellenőrzéséért, illetve Dr. Pósa Lászlót a minták legyártásáért és a szakmai konzultációkért.

Ezúton fejezem ki hálámat Juhász Zoltán tanár úrnak, hogy a gimnáziumi évek alatt felkeltette az érdeklődésemet a matematika és fizika tárgyak iránt. Kezdeti iránymutatása nélkül ezeket az eredményeket nem érhettem volna el. Végül pedig szívből köszönöm családomnak, hogy szeretettel és jó tanácsokkal fordulnak felém, illetve gondolatébresztő beszélgetésekkel kövezték ki a dolgozat megírásához vezető rögzös utat.

A. Hopfield hálózat energia-minimalizációs tulajdonsága

A Hopfield-hálózat energiája

$$E = -\frac{1}{2} \sum_{i,j} [\underline{W} \circ (\underline{x} \circ \underline{x})]_{ij} + \underline{\theta} \cdot \underline{x} = -\frac{1}{2} \sum_{i,j} w_{ij} \cdot x_i x_j + \sum_i \theta_i x_i$$

képlet szerint számolható. A következőkben bebizonyítom, hogy a frissítési szabályok figyelembevételével a hálózat energiája csak csökkenhet, vagy stagnál, vagyis a Hopfield-hálózat valóban minimalizálja energiafüggvényét.

Tegyük fel, hogy adott egy hálózat $\underline{W}$ súlymátrixszal és $\underline{x}$ neuronokkal. Nézzük meg, hogy a k . neuron értékét megváltoztatva

$$x_i^* - x_i = \Delta_k \cdot \delta_{ik}, \quad \text{ahol } \Delta_k \triangleq x_k^* - x_k$$

hogyan változik a hálózat energiája.

Az energiafüggvény két tag összege, írjuk fel tehát a megváltozását mint a két tag megváltozásának összegét

$$\Delta E = \Delta E_1 + \Delta E_2.$$

Behelyettesítés után

$$\Delta E = -\frac{1}{2} \sum_{i,j} w_{ij} (x_i^* x_j^* - x_i x_j) + \sum_i \theta_i (x_i^* - x_i).$$

$x_i^* - x_i$ definíciójának ismeretében a megváltozás második tagja

$$\Delta E_2 = \theta_k (x_k^* - x_k) = \theta_k \Delta_k$$

kifejezésre egyszerűsödik.

A második tag átalakításához látnunk kell, hogy a $\underline{W} \circ (\underline{x} \circ \underline{x})$ kifejezésben csak azon tagok értéke térhet el egymástól, amelyekben a k index szerepel. Az így kiválasztott sor és oszlop képletét helyettesítsük be és vegyük észre, hogy a megfelelő tag leválasztásával a kapott szummák értéke a megváltozott neuron értékétől függetlenek, sőt csillagozott és nem-csillagozott változókra is megegyeznek

$$\begin{aligned} \Delta E_1 &= -\frac{1}{2} \left(\sum_{i,j} w_{ij} (x_i^* x_j^* - x_i x_j) \right) \\ &= -\frac{1}{2} \left(2x_k^* \sum_{i \neq k} w_{ik} x_i^* + 2w_{kk} x_k^{*2} - \left(2x_k \sum_{i \neq k} w_{ik} x_i + 2w_{kk} x_k^2 \right) \right). \end{aligned}$$

Vezessük be az A_k jelölést

$$A_k \triangleq \sum_{i \neq k} w_{ik} x_i^* = \sum_{i \neq k} w_{ik} x_i$$

és alakítsuk tovább a képletet

$$\Delta E_1 = -(x_k^* A_k + w_{kk} x_k^{*2} - (x_k A_k + w_{kk} x_k^2)) = -(A_k (x_k^* - x_k) + w_{kk} (x_k^{*2} - x_k^2))$$

Mivel $\forall i : x_i \in \{-1, 1\}$, ezért $|x_k^*|^2 = |x_k|^2 = 1$, vagyis a megváltozásban szereplő első energiatag kifejezése még tovább egyszerűsödik

$$\Delta E_1 = -A_k \Delta_k.$$

Végeredményben tehát

$$\Delta E = -A_k \Delta_k + \theta_k \Delta_k = \Delta_k (\theta_k - A_k).$$

A frissítési szabályok értelmében pedig vizsgáljuk meg A_k és θ_k viszonyát

$$A_k = \sum_{i \neq k} w_{ik} x_i \geq \theta_k \longrightarrow x_k^* = 1 \implies \Delta_k \in \{0, 2\} \implies \Delta E \leq 0$$

$$A_k = \sum_{i \neq k} w_{ik} x_i < \theta_k \longrightarrow x_k^* = -1 \implies \Delta_k \in \{0, -2\} \implies \Delta E \leq 0$$

Vagyis a megfelelően megválasztott szabálynak és energia-függvénynek köszönhetően könnyen beláthattuk, hogy a Hopfield-hálózat minden lépésben csökkenti vagy minimalizálja energiafüggvényét.

B. Max-cut probléma mint optimalizációs feladat

A maximális vágás (*max-cut*) egy NP-nehéz feladat, melyben a $G(V, E)$, $V \in \{1, 2, \dots, N\}$ csúcsokkal és E élekkel rendelkező gráf olyan $S \subseteq V$ szerinti

$$\delta(S) = \{ij \in E : i \in S, j \in V\}$$

particionálását keressük, melyre a

$$w(\delta(S)) = \sum_{ij \in \delta(S)} w_{ij}$$

súly, vagyis súlyozatlan gráf esetén az S és $S \setminus V$ algráfok között futó élek száma, maximális.

Ennek kvadratikus egészértékű programozási feladattá alakításához a [17] forrásban bemutatott módszert használom. Vezessük be a $\underline{x} \in \{-1, 1\}^N$ vágás-vektort, mely

$$x_i = \begin{cases} 1 & i \in S \\ -1 & i \in S \setminus V \end{cases}$$

összefüggés szerint kódolja az adott particionálást. Ekkor a két algráf között futó élek száma

$$\sum_{ij \in \delta(S)} w_{ij} \frac{1 - x_i x_j}{2}$$

képlet szerint számolható. Itt ugyanis ij minden csúcson végigfut és ha kettő csúcs között fut él, de egy algráfba esnek, akkor a járuléku az összegben nulla, ha pedig külön algráfba kerültek, akkor $w_{ij} = 1$ következtében járuléku egységnyi. Végeredményben tehát az összefüggés valóban a maximalizálandó mennyiséget, az algráfok között futó élek számát adja súlyozatlan gráfra.

Alakítsuk át ezt a képletet, kihasználva, hogy $\underline{Q}$ szimmetrikus és $x_i x_i = 1$

$$\begin{aligned} \sum_{ij \in \delta(S)} q_{ij} \frac{1 - x_i x_j}{2} &= \frac{1}{2} \sum_{i < j} q_{ij} (1 - x_i x_j) \\ &= \frac{1}{4} \sum_{i, j} q_{ij} (1 - x_i x_j) \\ &= \sum_i \left(\sum_j q_{ij} x_i x_i - \sum_j x_i x_j \right). \end{aligned}$$

Bevezetve ekkor a q_{ij}^* súlyokat tartalmazó $\underline{Q}^*$ mátrixot, ahol

$$q_{ij}^* = \begin{cases} -q_{ij} & i \neq j \\ \sum_j q_{ij} - q_{ii} & i = j \end{cases}$$

behelyettesítés útján látható, hogy

$$\sum_{ij \in \delta(S)} q_{ij} \frac{1 - x_i x_j}{2} = \frac{1}{4} \underline{x}^T \underline{Q}^* \underline{x},$$

hiszen

$$\begin{aligned}
\frac{1}{4} \underline{x}^T \underline{Q}^* \underline{x} &= \sum_i x_i \sum_j q_{ij}^* x_j \\
&= \sum_i \sum_{i \neq j} (-q_{ij}) x_i x_j + \sum_i \left(\sum_j q_{ij} - q_{ii} \right) x_i x_i \\
&= - \sum_i \sum_{j \neq i} q_{ij} x_i x_j + \sum_{i,j} q_{ij} x_i x_i - \sum_{i,j} q_{ii} x_i x_i \\
&= \sum_{i,j} q_{ij} x_i x_i - \sum_{i,j} q_{ij} x_i x_j.
\end{aligned}$$

$\hat{\underline{Q}} = \frac{1}{2} Q^*$ bevezetésével pedig könnyen észrevehető, hogy a feladat igazából az

$$\underline{x} = \operatorname{argmax}_{\underline{x} \in \{\pm 1\}^N} \frac{1}{2} \underline{x}^T \hat{\underline{Q}} \underline{x} \longrightarrow \underline{x} = \operatorname{argmin}_{\underline{x} \in \{\pm 1\}^N} - \frac{1}{2} \underline{x}^T \hat{\underline{Q}} \underline{x}$$

egyenlet megoldása, ami épp a Hopfield-hálózat energiafüggvényének alternatív alakja.

Hivatkozások

- [1] J. J. Yang, D. B. Strukov, and D. R. Stewart. *Memristive devices for computing*. Nature Nanotechnology, **8**, 13 (2013).
- [2] M. Le Gallo and A. Sebastian. *An overview of phase-change memory device physics*. Journal of Physics D: Applied Physics, **53**, 213002 (2020).
- [3] L. Pósa, M. Abbassi, P. Makk, B. Santa, C. Nef, M. Csontos, M. Calame, and A. Halbritter. *Multiple physical timescales and dead time rule in few-nm sized graphene-SiO_x -graphene memristors*. Nano Letters, **17**, 6783 (2017).
- [4] F. Cai, S. Kumar, T. Van Vaerenbergh, X. Sheng, R. Liu, C. Li, Z. Liu, M. Foltin, S. Yu, Q. Xia, J. J. Yang, R. Beausoleil, W. D. Lu, and J. P. Strachan. *Power-efficient combinatorial optimization using intrinsic noise in memristor Hopfield neural networks*. Nature Electronics, **3**, 409 (2020).
- [5] L. Pósa. *Resistive switching in ultrasmall nanogap devices*. Ph.D. thesis, Budapesti Műszaki és Gazdaságtudományi Egyetem (2019).
- [6] V. Karpov, Y. Kryukov, M. Mitra, and I. Karpov. *Crystal nucleation in glasses of phase change memory*. Journal of Applied Physics, **104**, 054507 (2008).
- [7] J. G. Fehérvári. *Nanoméretű fázisváltó memóriák időskáláinak kísérleti vizsgálata* (2020). TDK Dolgozat, Budapesti Műszaki és Gazdaságtudományi Egyetem.
- [8] K. Kohary and C. D. Wright. *Electric field induced crystallization in phase-change materials for memory applications*. Applied Physics Letters, **98**, 223102 (2011).
- [9] M. A. Zidan, J. P. Strachan, and W. D. Lu. *The future of electronics based on memristive systems*. Nature Electronics, **1**, 22 (2018).
- [10] Q. Xia and J. J. Yang. *Memristive crossbar arrays for brain-inspired computing*. Nature Materials, **18**, 309 (2019).
- [11] S. Ambrogio, P. Narayanan, H. Tsai, R. Shelby, I. Boybat, C. Nolfo, S. Sidler, M. Giordano, M. Bordini, N. Farinha, B. Killeen, C. Cheng, Y. Jaoudi, and G. Burr. *Equivalent-accuracy accelerated neural-network training using analogue memory*. Nature, **558**, 60 (2018).
- [12] C. Liu, M. Hu, J. P. Strachan, and H. H. Li. *Rescuing Memristor-Based Neuromorphic Design with High Defects*. In *Proceedings of the 54th Annual Design Automation Conference 2017*, DAC '17. Association for Computing Machinery, New York, NY, USA (2017).
- [13] R. Rutenbar. *Simulated annealing algorithms: an overview*. IEEE Circuits and Devices Magazine, **5**, 19 (1989).
- [14] D. Ielmini, F. Nardi, and C. Cagli. *Resistance-Dependent Amplitude of Random Telegraph Signal Noise in Resistive Switching Memories*. Applied Physics Letters, **96**, 053503 (2010).

- [15] M. R. Mahmoodi, H. Nili, and D. B. Strukov. *RX-PUF: Low Power, Dense, Reliable, and Resilient Physically Unclonable Functions Based on Analog Passive RRAM Crossbar Arrays*. In *2018 IEEE Symposium on VLSI Technology*, 99–100 (2018).
- [16] P. Balázs. *SiO₂ rezisztív kapcsolók vezetési mechanizmusának vizsgálata zajjelenségekkel keresztül* (2020). TDK Dolgozat, Budapesti Műszaki és Gazdaságtudományi Egyetem.
- [17] Z. J. Wu LY, Zhang XS. *Application of discrete Hopfield-type neural network for max-cut problem*. In *ICONIP*, 1439–44 (2001).