Negative capacitance and hyperdimensional computing for unconventional low-power computing

by

Justin C. Wong

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Sayeef Salahuddin, Chair
Professor Jan M. Rabaey
Professor Bruno A. Olshausen

Fall 2018

Negative capacitance and hyperdimensional computing for unconventional low-power computing

Abstract

Negative capacitance and hyperdimensional computing for unconventional low-power computing

by

Justin C. Wong

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Sayeef Salahuddin, Chair

Properties that emerge from the collective behavior of constituents at different length scales can be exploited to reduce power consumption below conventional limits in computing. At the device level, ferroelectric-dielectric coupling ("negative capacitance") can reduce energy consumption below $^1/_2\,CV^2$ in capacitors. However, this effect is still not well understood. We construct a microscopic model and analyze energy flow from the perspective of Poynting's theorem to clear up these misunderstandings. At the circuit level, high-dimensional distributed representations relax requirements on signal-to-noise ratio and supply voltage, and enable new architecture designs. Computing with these representations ("hyperdimensional computing") is natural for performing energy efficient cognitive computing at the application level. However, data in practice is always measured in some sort of representation, which may not be natural for hyperdimensional computing. We bridge this gap by proposing to use an approximation of the bispectrum to map data measured in practice into high-dimensional distributed representations for use with hyperdimensional computing.

To my mom

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.5 billion smartphones were sold in 2017 [1]. This is in addition to the more than 2.5 billion smartphones estimated to already be in use in the world [2]. This number is still growing and is projected to continue growing as smartphone penetration increases in countries around the world [3]. Add in the more than 17 billion connected devices from the growing internet of things (IoT) [4], and we have a rapidly expanding ecosystem of technology that is increasingly reliant on low power electronics. In order to support this ecosystem and allow it to continue to flourish, our underlying hardware needs to become more energy efficient.

To further complicate matters, the recent successes of deep artificial neural networks have sparked a "deep learning revolution" [5]. A wide range of industries are now racing to adopt deep learning in their technologies [6, 7, 8] including everyday technologies such as mobile applications and the IoT [9, 10, 11, 12]. However, training deep networks requires copious amounts of training data and time and is computationally expensive due to the need to frequently update potentially billions of weights over numerous training epochs [13, 14]. Furthermore, training a network to complete a *new* task typically requires retraining multiple layers—or even an entirely new neural network from scratch in many cases—on new training data [14, 15]. Clearly, this approach is energy inefficient and computationally infeasible for embedded systems. In order to solve this energy problem, we need innovations at every level of the computing stack.

At the device level, energy consumption cannot be reduced in conventional metal-oxide-semiconductor field-effect transistors (MOSFETs) due to thermal limits arising from Boltzmann statistics [16, 17]. Salahuddin and Datta proposed to overcome this limit in 2008 by using ferroelectric negative capacitance [18]. However, this effect is still not well understood. Some believe it is simply a theoretical artifact from an unphysical model and cannot exist. Even with the growing experimental evidence [19, 20, 21, 22, 23] for ferroelectric negative capacitance, there are still conceptual shortcomings surrounding the existence of negative capacitance, its stabilization, and whether or not it would actually provide any energy savings if it could work.

If we did succeed in overcoming the thermal barrier at the device level, there are still signal-to-noise ratio (SNR) requirements at the circuit level needed to maintain logical con-

sistency, especially in memory [24, 25, 17]. Clever circuit design can improve noise immunity by rejecting internal noise sources to an extent [25]. However, there are still noise sources such as device variation and crosstalk that worsen with scale and increased integration density [25, 26]. High-dimensional distributed representations can improve noise immunity and are natural for energy efficient cognitive computing ("hyperdimensional computing") [27, 28, 29]. Hyperdimensional computing can potentially perform more human-like inference [27] with less data and in one shot compared to conventional machine learning methods [28, 29]. However, the mathematics of using hyperdimensional computing to solve major learning problems are still not well understood.

In this thesis, we address these problems in three chapters:

- Chapter 2 fills in the theoretical gaps of ferroelectric negative capacitance. We first construct a simple but physically intuitive microscopic model to illustrate how negative capacitance arises and can be stabilized in a prototypical displacive ferroelectric. We show that the phenomenological model originally used to derive ferroelectric negative capacitance can have correspondences with physical crystal configurations. Then, we analyze energy flow in a ferroelectric-dielectric system from the perspective of Poynting's theorem. We show that overall energy dissipation is reduced below the $^1/_2\,CV^2$ limit during charging and discharging.

- In Chapter 3, we discuss the notion of information representation and its implications in computation. In the context of automata theory and formal languages, we illustrate how distributed representations can be more computationally efficient than local representations. High-dimensional distributed representations possess even greater potential computational efficiency and are robust against errors due to redundancy.

- Chapter 4 proposes a method for bridging the gap between data measured in practice and high-dimensional distributed representations. We propose to use an approximation of the bispectrum to encode data measured in practice into high-dimensional distributed representations for use in hyperdimensional computing. We then apply our method to a small dataset of music to demonstrate one-shot unsupervised learning.

Finally, we end in Chapter 5 with a brief summary of the key results and discuss future work.

Chapter 2

Negative Capacitance

After the publication of constant-field scaling rules by Dennard et al. in 1974 [30], the semiconductor industry raced to reap the benefits of *Dennard scaling* in metal-oxide-semiconductor field-effect transistors (MOSFETs). By reducing device dimensions, supply voltages could also be reduced while preserving electric field intensities inside MOSFETs to maintain constant power density [30, 31, 16, 17]. The net result was faster devices, higher circuit densities, and lower power dissipation per circuit [30, 31, 16, 17]. Unfortunately, constant-field scaling rules could only be followed approximately since not all device factors scale with geometry. Temperature and material energy band gap are two primary nonscaling factors that result in nonscaling threshold voltage and built-in potential [16, 17]. Consequently, supply voltages could not be scaled as aggressively as device dimensions, resulting in increased electric fields and short-channel effects. To mitigate the worsening of these effects with scale, Baccarani et al. defined generalized scaling rules that preserve electric field profiles and potential profiles [32]. However, this did not work without limit since electric field intensities still increased. Nevertheless, this golden era of geometrical scaling continued for decades from the 1970s until around 2006 when leakage effects began to dominate [33].

As geometrical scaling came to an end, a new era of *equivalent scaling* began [34, 33]. By changing certain device characteristics, transistor dimensions could be equivalently scaled instead of physically scaled. Metal gates and high-κ dielectrics enabled equivalently thinner oxides without needing to physical decrease oxide thickness [35]. The use of germanium and strained silicon increased carrier mobility, resulting in faster devices [36, 37, 38, 39, 40, 35]. Non-planar structures such as the FinFET enabled better gate control over the semiconductor channel to mitigate short-channel effects [41, 42]. While these innovations helped devices continue to scale, they still did not change the fundamental limitations imposed by the nonscaling factors mentioned above. Consequently, equivalently scaling only brought devices closer to their fundamental physics limitations. Nonscaling temperature, for example, still places a lower limit on subthreshold swing due to Boltzmann statistics.

To overcome this Boltzmann limit, Salahuddin and Datta proposed in 2008 to replace the gate oxide with a ferroelectric insulator exhibiting negative capacitance [18]. This would, in theory, reduce the body factor below the ideal limit of one, resulting in an effective am-

plification of the gate voltage at the semiconductor channel and steeper subthreshold slope. This idea was exciting [43, 44] but was not without controversy: (1) in their original work, Salahuddin and Datta used a phenomenological model of ferroelectricity, which has no direct correspondence with physical configurations, leading some to believe that ferroelectric negative capacitance is an artifact of the phenomenological model; (2) even if ferroelectric negative capacitance does exist, it must be unstable and cannot be stabilized without making the capacitance positive or breaking the ferroelectric into domains; and (3) the use of negative capacitance does not reduce net energy cost since there must be an active source that supplies extra energy. These controversies have somewhat subsided over the past decade due to growing experimental evidence [19, 20, 21, 22, 23] of ferroelectric negative capacitance. However, there are still conceptual gaps and misunderstandings regarding the origin of ferroelectric negative capacitance and the above main criticisms.

In this chapter, we demystify ferroelectric negative capacitance and address the above criticisms. First, we construct a simple and physically intuitive microscopic model of ferroelectricity to demonstrate how negative capacitance can arise and be stabilized in a physical configuration of ions. We provide a correspondence between the microscopic model and phenomenological models to alleviate concerns about ferroelectric negative capacitance artificially manifesting from the use of a phenomenological model. Finally, we analyze energy flow from the perspective of Poynting's theorem during charging and discharging of a ferroelectric-dielectric capacitor to show that energy flows directly from the ferroelectric into the dielectric during charging and vice versa during discharging. Thus, net energy dissipation is reduced because no active source is needed to supply extra energy.

2.1 Introduction

2.1.1 Subthreshold Swing and the Boltzmann Limit

Recall from solid state physics that the density of electron states in the n-th energy band of an ideal crystal is given by [45, 46, 47]

$$D_n(\varepsilon) = \int_{S_n(\varepsilon)} \frac{dS}{4\pi^3} \frac{1}{|\nabla \varepsilon_n(\mathbf{k})|} \tag{2.1}$$

where $S_n(\varepsilon)$ is the surface of constant energy ε in k-space corresponding to the n-th energy band; and $\varepsilon_n(\mathbf{k})$ is the dispersion relationship. Assuming the crystal is a semiconductor, we can perform a Taylor series expansion of the dispersion relationship about a critical point $\mathbf{k}_0$ corresponding to the conduction or valence band edge:

$$\varepsilon_n(\mathbf{k}) = \varepsilon_n(\mathbf{k}_0) + [(\mathbf{k}-\mathbf{k}_0)\cdot\nabla]\varepsilon_n(\mathbf{k}_0) + \frac{1}{2}[(\mathbf{k}-\mathbf{k}_0)\cdot\nabla]^2\varepsilon_n(\mathbf{k}_0) + \cdots \tag{2.2}$$

Note that the linear terms go to zero since we are expanding about a critical point, and the quadratic terms can be diagonalized via a transformation to principal axes [46]. Truncating this expansion at second-order results in the usual parabolic approximation where the

curvature is characterized by an effective mass tensor. This results in ellipsoidal surfaces of constant energy and a density of states with the following trend:

$$D_{n\mathbf{k}_0}(\varepsilon) \sim (\varepsilon - \varepsilon_n(\mathbf{k}_0))^\eta \tag{2.3}$$

η is a constant factor that depends on the dimensionality of the material (e.g. $\eta = 1/2$ for a three-dimensional crystal). To account for degeneracy, we sum the density of states over all energy bands and corresponding critical $\mathbf{k}$-points:

$$D(\varepsilon) = \sum_{n,\mathbf{k}_0} D_{n\mathbf{k}_0}(\varepsilon) \sim (\varepsilon - \varepsilon_0)^\eta \tag{2.4}$$

Note that we have now replaced the degenerate energy level $\varepsilon_n(\mathbf{k}_0)$ with simply ε_0.

The thermal average occupancy of these states is dictated by the Fermi-Dirac distribution [48]

$$f(\varepsilon) = \frac{1}{1 + \exp\left(\frac{\varepsilon - \mu}{k_\mathrm{B}T}\right)} \tag{2.5}$$

where μ is the chemical potential or Fermi level; k_B is the Boltzmann constant; and T is temperature. In the classical concentration limit (e.g. in a non-degenerately doped material), the Fermi-Dirac distribution approaches the Boltzmann distribution [48]:

$$f(\varepsilon) \approx \exp\left(-\frac{\varepsilon - \mu}{k_\mathrm{B}T}\right) \tag{2.6}$$

Thus, the distribution of electrons as a function of energy is given by

$$dn = f(\varepsilon)D(\varepsilon)\,d\varepsilon \sim (\varepsilon - \varepsilon_0)^\eta \exp\left(-\frac{\varepsilon - \mu}{k_\mathrm{B}T}\right) d\varepsilon \tag{2.7}$$

Notice that Boltzmann statistics dominates the density of states, resulting in an overall exponentially decaying distribution of electrons with increasing energy. This means that if we can *decrease* the electron energy relative to the chemical potential, then we can exponentially *increase* the number of electrons. We can accomplish this by applying an appropriate electric field that establishes a surface potential ψ_s in the semiconductor:

$$dn \sim [(\varepsilon - q\psi_\mathrm{s}) - (\varepsilon_0 - q\psi_\mathrm{s})]^\eta \exp\left[-\frac{(\varepsilon - q\psi_\mathrm{s}) - \mu}{k_\mathrm{B}T}\right] d\varepsilon \tag{2.8}$$

$$\sim (\varepsilon - \varepsilon_0)^\eta \exp\left(-\frac{\varepsilon - \mu}{k_\mathrm{B}T}\right) \exp\left(\frac{q\psi_\mathrm{s}}{k_\mathrm{B}T}\right) d\varepsilon \tag{2.9}$$

This is the basis of the *field-effect transistor*. Although the above analysis focused on electron states, the results apply equally to holes.

In a *metal-oxide-semiconductor* field-effect transistor (MOSFET), this mechanism is used to modulate the conductivity of a semiconducting *channel* to turn the channel "on" or "off".

The modulating electric field is controlled by a gate voltage V_g applied at a gate terminal on the other side of an oxide sitting on the semiconductor. Simultaneously, an additional lateral electric field may be applied across the length of the channel to induce a current across the channel. When the gate voltage is above a certain *threshold voltage* V_T, the number of charge carriers in the channel—and, in turn, the channel conductivity—will be high enough for substantial current to flow. Reducing the gate voltage below the threshold voltage will, ideally, lower the channel conductivity enough to stop the current flow. In other words, the *subthreshold current* is ideally zero. Unfortunately, this does not happen due to Boltzmann statistics as explained next.

From (2.9) and the discussion above, the number of charge carriers in the channel increases exponentially with the semiconductor surface potential. Consequently, the subthreshold current also has an exponential dependence on the surface potential:

$$I_\mathrm{s} \sim \exp\left(\frac{q\psi_\mathrm{s}}{k_\mathrm{B}T}\right) \tag{2.10}$$

Thus, the change in gate voltage needed to change the current by an order of magnitude is given by [49, 31, 16, 17]

$$S = \frac{\partial V_\mathrm{g}}{\partial \log_{10}(I_\mathrm{s})} = \frac{\partial V_\mathrm{g}}{\partial \psi_\mathrm{s}} \frac{\partial \psi_\mathrm{s}}{\partial \log_{10}(I_\mathrm{s})} = \underbrace{\frac{k_\mathrm{B}T}{q}\ln(10)}_{\substack{\approx 60 \text{ mV/decade} \\ \text{at } T=300 \text{ K}}} \underbrace{\left(1 + \frac{C_\mathrm{s}}{C_\mathrm{ins}}\right)}_{m} \tag{2.11}$$

This is called *subthreshold swing*. The first underbraced term manifests from the underlying Boltzmann statistics (as evident by the presence of the Boltzmann constant and temperature) and is known as the "Boltzmann limit" on subthreshold swing. At room temperature ($T \approx 300$ K), the Boltzmann limit is ≈ 60 mV/decade. The second underbraced term m is called the *body factor*, which characterizes the capacitive coupling $\partial V_\mathrm{g}/\partial \psi_\mathrm{s}$ between the gate terminal and the semiconductor surface. If the gate has perfect control over the channel, then $m = 1$ and the subthreshold swing is only limited by Boltzmann statistics. Generally, however, there is capacitive coupling between the semiconductor surface and other parts of the device, resulting in a semiconductor capacitance C_s. Thus, the gate insulator C_ins and semiconductor capacitance C_s form a capacitive voltage divider that increases the body factor (i.e. $m > 1$) *assuming positive capacitances* and, consequently, increases subthreshold swing.

This is undesirable because an ideal switch should have zero subthreshold swing in order to abruptly switch with infinitesimally small changes in gate voltage. From (2.11), we see that the subthreshold swing can be minimized by reducing the Boltzmann limit and the body factor. The Boltzmann limit, however, cannot be changed without expending additional energy to cool below room temperature. Alternatively, the Boltzmann limit can be circumvented by using other devices that simply do not rely on Boltzmann statistics. One popular example is the tunnel field-effect transistor (TFET), but it comes with its own practical challenges [50, 51]. Instead, the semiconductor industry has traditionally focused on

decreasing the body factor as close as possible to the presumed ideal limit of $m = 1$ by maximizing gate control. Geometrical scaling and equivalent scaling (e.g. high-κ dielectrics, metal gates, and non-planar multigate structures) have directly improved gate control by decreasing the equivalent oxide thickness or increasing the gate area around the channel to increase C_{ins} [35, 41, 42]. Other technologies, such as fully depleted silicon-on-insulator (FD-SOI) technology, indirectly improve gate control by instead reducing the semiconductor capacitance C_{s} [52]. These examples are just some of the innovations that have allowed the semiconductor industry to improve gate control close to its presumed ideal limit of perfect capacitive coupling, leaving subthreshold swing practically limited by just the Boltzmann limit. Thus, it would seem that there are no further improvements that can be made aside from cooling below room temperature.

2.1.2 Negative Capacitance

In 2008, Salahuddin and Datta proposed to use *negative capacitance* to overcome the Boltzmann limit [18]. By replacing the gate insulator with a material that exhibits negative capacitance, the body factor can, in principle, be decreased below its presumed ideal limit. In other words, instead of striving for perfect capacitive coupling ($m = 1$), we seek to effectively amplify the gate voltage at the channel ($m < 1$). This idea is fairly straightforward if we simply examine (2.11). However, it is not immediately obvious what negative capacitance means or how we can find a material system that exhibits such a property.

In a general sense, negative capacitance can be understood in terms of positive feedback [47, 18, 53]. Following Salahuddin and Datta's derivation [18], if we have an ordinary linear capacitor C, and we apply a voltage v, then we expect the capacitor to store charge $q = Cv$. However, if there is a positive feedback effect, then the charge will amplify the input voltage $v \to v + \alpha_{\text{f}} q$ where α_{f} characterizes the strength of the feedback. This results in more charge $q = C(v + \alpha_{\text{f}} q)$ stored in the capacitor. Solving for charge in terms of voltage, we obtain an expression for the effective capacitance:

$$C_{\text{ins}} = \frac{C}{1 - \alpha_{\text{f}} C} \tag{2.12}$$

Notice that the effective capacitance becomes *negative* if the feedback is strong enough (i.e. $\alpha_{\text{f}} C > 1$), but the system becomes unstable. We can stabilize the negative capacitance by placing a positive capacitor C_{s} in series. This results in an equivalent capacitance for the combined system:

$$C_{\text{eq}} = \frac{C_{\text{ins}} C_{\text{s}}}{C_{\text{ins}} + C_{\text{s}}} = \frac{|C_{\text{ins}}||C_{\text{s}}|}{|C_{\text{ins}}| - C_{\text{s}}} \tag{2.13}$$

If the positive capacitor is thick enough (i.e. $C_{\text{s}} < |C_{\text{ins}}|$), then the overall capacitance remains positive, and the system is stable even though the insulator capacitance is still negative.

This analysis implies that any charge-based system with positive feedback may exhibit negative capacitance. In 2014, I published a paper exploring the possibility of achieving

negative capacitance with an electromechanical positive feedback effect [54] and wrote a Master's report on it in 2015 [55]. I found that electromechanical negative capacitance is possible in principle but difficult to achieve in materials commonly used in electronic devices. In contrast, Salahuddin and Datta proposed to use ferroelectrics as their material system. In order to understand how ferroelectrics may exhibit negative capacitance, we must first discuss Landau theory.

2.1.3 Landau Theory

Ferroelectrics are most commonly modelled using a phenomenological theory known as *Landau theory*. Landau theory takes into account transformations of crystal symmetry in order to model a Landau free energy density $f_L(\eta, T)$ as a function of an *order parameter* η and temperature T [56, 57]. The order parameter is chosen to reflect the symmetry transformations of interest. For example, polarization and magnetization are the typically chosen order parameters for ferroelectrics and ferromagnets respectively.

By taking into account known transformations of crystal symmetry, the Landau free energy density can be expanded into a series that reflects the same symmetries. For example, if we consider a one-dimensional order parameter that displays even symmetry under a phase transition as the temperature varies, then the Landau free energy density has the form

$$f_{\mathrm{L}}(\eta, T) = f_0 + \frac{1}{2}\alpha(T)\eta^2 + \frac{1}{4}\beta(T)\eta^4 + \frac{1}{6}\gamma(T)\eta^6 + \cdots \tag{2.14}$$

where $\alpha(T)$, $\beta(T)$, and $\gamma(T)$ are Landau free energy coefficients. Notice that the model has no direct correspondences with physical configurations of the crystal, but it often ends up working well in practice because it approximates the "true" free energy density, which is normally calculated from first principles or extracted from experiments [57]:

$$f_{\mathrm{L}} \approx f = u - Ts \tag{2.15}$$

u is the internal energy density, and s is the entropy per unit volume. The exact differential is

$$\begin{aligned} df &= du - d(Ts) \\ &= (T\,ds + \zeta\,d\eta) - (T\,ds + s\,dT) \\ &= -s\,dT + \zeta\,d\eta \end{aligned} \tag{2.16}$$

where ζ and η are generalized force and displacement respectively and are thermodynamic conjugate variables. Thus, the free energy density is minimized under small fluctuations in temperature and order parameter:

$$(\delta f)_{T,\eta} = 0 \tag{2.17}$$

The generalized force is determined by

$$\zeta = \left(\frac{\partial f}{\partial \eta}\right)_T \tag{2.18}$$

Similar expressions can be found for the entropy per unit volume as well as the usual Maxwell relations.

If we now consider a ferroelectric with one-dimensional order parameter P for polarization and generalized force E for electric field, then from (2.14) the Landau free energy density is

$$f_{\mathrm{L}}(P,T) = f_0 + \frac{1}{2}\alpha_0(T-T_0)P^2 + \frac{1}{4}\beta P^4 + \frac{1}{6}\gamma P^6 \tag{2.19}$$

Notice that we have truncated the series at sixth order, and we have assumed a continuous phase transition where the second-order Landau free energy coefficient $\alpha(T)$ is linear in temperature for simplicity (α_0 is positive, and T_0 is the phase transition temperature). In order for the ferroelectric to have overall stability, γ must be positive. We also assume $\beta > 0$ for simplicity. From (2.18) and (2.15), the electric field is

$$E = \left(\frac{\partial f}{\partial P}\right)_T \approx \left(\frac{\partial f_{\mathrm{L}}}{\partial P}\right)_T = \alpha_0(T-T_0)P + \beta P^3 + \gamma P^5 \tag{2.20}$$

If we now plot polarization as a function of electric field, then we obtain the ferroelectric S curve, which has a region of negative slope. This is the negative capacitance region. If we differentiate the electric field with respect to polarization, then we can calculate the inverse electric susceptibility:

$$\frac{1}{\epsilon_0 \chi_{\mathrm{e}}} = \left(\frac{\partial E}{\partial P}\right)_T = \alpha_0(T-T_0) + 3\beta P^2 + 5\gamma P^4 \tag{2.21}$$

ϵ_0 is vacuum permittivity. For small polarizations, the inverse electric susceptibility is

$$\frac{1}{\epsilon_0 \chi_{\mathrm{e}}} \approx \alpha_0(T-T_0) \tag{2.22}$$

Thus, if we cool the ferroelectric below the critical temperature, then the electric susceptibility becomes negative at small polarizations! This is the basis of ferroelectric negative capacitance. In general, however, ferroelectrics are unstable near zero polarization. Thus, it is conventionally believed that (2.22) is never applicable at equilibrium and that we must instead use (2.21), which predicts a positive susceptibility at equilibrium.

2.2 Microscopic Origin of the S Curve

The conventional analysis of a negative capacitance transistor uses a top-down approach that relies heavily on a phenomenological model of ferroelectricity: a Landau free energy function is expanded about an order parameter into a mathematical series that takes into account known transformations of the crystal symmetry. The coefficients of this series expansion are then fit a posteriori to experimental data. While this phenomenological approach often works well enough in practice, it lacks any physical intuition at the microscopic scale. This

results in an uncomfortable disconnect: ferroelectric negative capacitance requires the existence of an unstable region in a theoretical S curve P-E loop, but this S curve is derived from a phenomenological energy that has no direct correspondences with any microscopic configurations. This has led some to view the S curve as merely an artifact or theoretical construct that does not exist in real ferroelectrics. In this section, we resolve this confusion by using a bottom-up approach from simple microscopics to illustrate how the ferroelectric S curve can correspond to physical configurations in a real crystal and is not simply theoretical.

We start by considering the ions of an ideal three dimensional ferroelectric crystal. We use the adiabatic principle to model an effective ionic potential energy that includes the electronic polarizabilities of the ions. We then separate this potential energy into local effective potential energies $V_i(\mathbf{u}_i)$ for each ion where $\mathbf{u}_i$ is the relative displacement of the i-th ion from its surrounding valence electrons. Thus, each ion experiences a local effective interatomic force described by

$$\mathbf{F}_{\text{ion}}^{(i)} = -\nabla V_i \tag{2.23}$$

The ions also interact with each other nonlocally, which we model using Coulomb forces (we neglect lattice distortions for simplicity). As the ions vibrate, they form pseudodipoles $\mathbf{p}_i$ through ion displacement and electronic polarization. Each pseudodipole produces an electric field at a position $\mathbf{r}$ given by

$$\mathbf{E}_{\text{dipole}}^{(i)}(\mathbf{r}) = \frac{3[\mathbf{p}_i \cdot (\mathbf{r} - \mathbf{R}_i)](\mathbf{r} - \mathbf{R}_i) - \mathbf{p}_i \left|\mathbf{r} - \mathbf{R}_i\right|^2}{4\pi\epsilon_0 \left|\mathbf{r} - \mathbf{R}_i\right|^5} \tag{2.24}$$

where $\mathbf{R}_i$ is the idealized equilibrium lattice site of the i-th ion. Thus, there is a nontrivial superposition of short-range and long-range Coulomb forces that results in a collective dipole field

$$\mathbf{E}_p(\mathbf{r}) = \sum_i \mathbf{E}_{\text{dipole}}^{(i)}(\mathbf{r}) \tag{2.25}$$

where the summation iterates over all ions in the crystal. We expect this field to dominate local interatomic forces for certain crystal configurations, resulting in a lattice instability that leads to a ferroelectric phase transition.

Next we use (2.24) and (2.25) to calculate the collective dipole field. The summation in (2.25) is generally nontrivial and is conditionally convergent for infinite crystals. However, if we use a mean field approximation and replace the individual dipole moments with their corresponding averages, then the summation can be decomposed into simpler sums over different interpenetrating sublattices that correspond to different ions. These lattice sums can be evaluated using a variety of methods, and we provide references to some in [58, 59]. The resultant dipole field evaluated at a B ion is then given approximately by

$$\begin{aligned}\mathbf{E}_p(\mathbf{R}_{\text{B}}) \approx{}& \frac{\langle \mathbf{p}_{\text{A}} \rangle}{4\pi\epsilon_0 v_{\text{cell}}} \left(-4\pi + \frac{4\pi}{3}\right) + \frac{\langle \mathbf{p}_{\text{B}} \rangle}{4\pi\epsilon_0 v_{\text{cell}}} \left(-4\pi + \frac{4\pi}{3}\right) \\ &+ \frac{\langle \mathbf{p}_{\text{O}_\parallel} \rangle}{4\pi\epsilon_0 v_{\text{cell}}} \left(-4\pi + \frac{4\pi}{3} + 30\right) + \frac{2\langle \mathbf{p}_{\text{O}_\perp} \rangle}{4\pi\epsilon_0 v_{\text{cell}}} \left(-4\pi + \frac{4\pi}{3} - 15\right)\end{aligned} \tag{2.26}$$

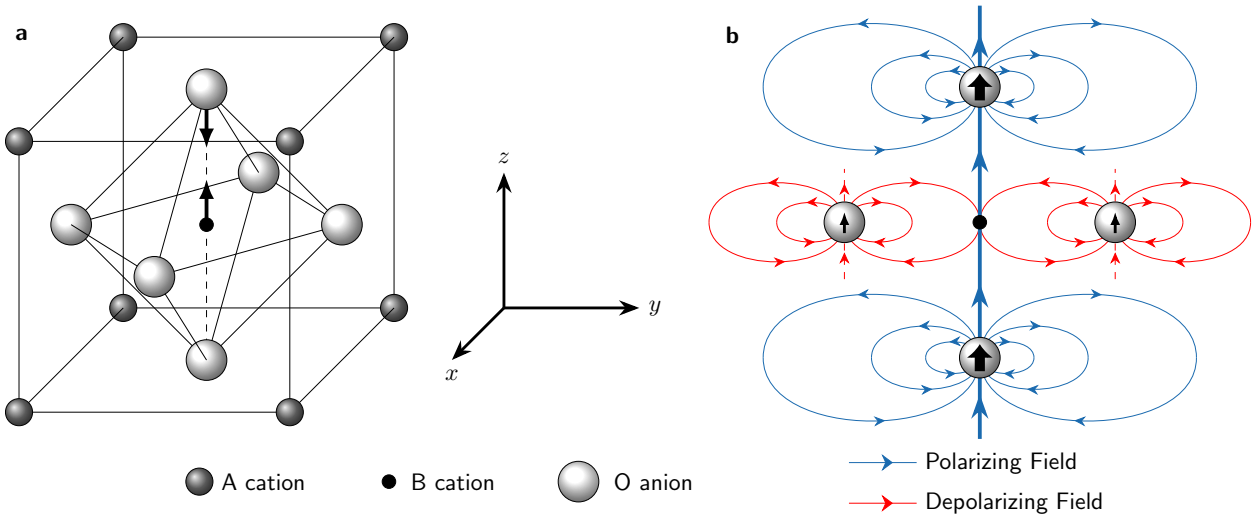

Figure 2.1: (a) Conventional unit cell of a ferroelectric perovskite (ABO_3). There are eight A cations (dark gray), one B cation (black), and six O anions (white). We divide the O ions into two groups: the two $O_\parallel$ ions at the ends of the dashed vertical line, which denotes the line of B ion displacement; and the remaining four $O_\perp$ ions. For simplicity, we assume that all relative displacements are along the z-axis. (b) Schematic of the dipole fields in the (200) plane. The $O_\parallel$ ions produce polarizing fields (blue) at the B ion while the $O_\perp$ ions produce depolarizing fields (red). The $O_\parallel$ ions are more polarized, so their fields dominate. The result is an overall polarizing field at the B ion.

where each parenthesized sum $(-4\pi + \cdots)$ is the result of a corresponding sublattice sum; $\langle \mathbf{p}_A \rangle$, $\langle \mathbf{p}_B \rangle$, $\langle \mathbf{p}_{O_\parallel} \rangle$, and $\langle \mathbf{p}_{O_\perp} \rangle$ are the mean dipole moments of the A, B, $O_\parallel$, and $O_\perp$ ions respectively; and v_{cell} is the volume of a unit cell. The $O_\parallel$ ions are the O ions that lie along the line of B ion displacement while the $O_\perp$ ions are the remaining O ions (see Fig. 2.1a). To further complicate matters, the dipole field in (2.26) is dependent on the mean ion dipole moments, which are in turn dependent on the dipole field—a self-consistent solution is needed. These calculations are beyond the scope of this section (Slater has performed such a calculation for $BaTiO_3$ in [60]), but we provide the qualitative result here: the polarization of a ferroelectric perovskite is typically dominated by the ion displacement of the B cations and by the electronic polarization of the $O_\parallel$ anions. Thus, we assume that the remaining ions have $\langle \mathbf{p}_A \rangle \approx \langle \mathbf{p}_{O_\perp} \rangle \approx 0$, and the dipole field simplifies to

$$\mathbf{E}_p(\mathbf{R}_B) \approx \frac{\langle \mathbf{p}_B \rangle}{4\pi\epsilon_0 v_{\text{cell}}} \left(-4\pi + \frac{4\pi}{3} \right) + \frac{\langle \mathbf{p}_{O_\parallel} \rangle}{4\pi\epsilon_0 v_{\text{cell}}} \left(-4\pi + \frac{4\pi}{3} + 30 \right) \tag{2.27}$$

If we also assume that $\langle \mathbf{p}_B \rangle$ and $\langle \mathbf{p}_{O_\parallel} \rangle$ contribute approximately equally to the polarization (i.e. $\langle \mathbf{p}_B \rangle \approx \langle \mathbf{p}_{O_\parallel} \rangle$), then we can write the dipole field in terms of macroscopic polarization

P:

$$\mathbf{E}_p(\mathbf{R}_\mathrm{B}) \approx -\frac{\mathbf{P}}{\epsilon_0} + \frac{\mathbf{P}}{3\epsilon_0} + \frac{15\mathbf{P}}{4\pi\epsilon_0} \tag{2.28}$$

Notice that $-\mathbf{P}/\epsilon_0$ is simply the macroscopic electric field that arises due to bound charge at the surface of a dielectric and is the depolarization field that acts to suppress polarization in typical dielectrics. $\mathbf{P}/3\epsilon_0$ is the standard Lorentz correction due to treating the dipoles as a continuum in a spherical cavity. These two terms combine into $-2\mathbf{P}/3\epsilon_0$, which is the usual dipole field in the Clausius-Mossotti picture. However, in the case of our perovskite ferroelectric, the O anions contribute a large positive term $\sim 15\mathbf{P}/4\pi\epsilon_0$, which makes the overall dipole field positive in the direction of polarization at the B cations. We can understand it better by examining the schematic shown in Fig. 2.1b where we have sketched the dipole fields of the $\mathrm{O}_\parallel$ and $\mathrm{O}_\perp$ ions in the (200) plane. The $\mathrm{O}_\parallel$ ions produce polarizing fields at the B ions while the $\mathrm{O}_\perp$ ions produce depolarizing fields. However, $|\langle\mathbf{p}_{\mathrm{O}_\parallel}\rangle| \gg |\langle\mathbf{p}_{\mathrm{O}_\perp}\rangle|$ as discussed above, so the $\mathrm{O}_\parallel$ dipole fields end up dominating. Thus, there is an overall polarizing field that tends to spontaneously polarize the B ions, which in turn help polarize the $\mathrm{O}_\parallel$ ions.

In addition to the contributions from polarization, we also need to account for the local effective potential energies. For simplicity, we assume all relative displacements are along the z-axis and are denoted by z_i. Thus, we can perform a simple series expansion of the local effective potential energies

$$V_i(z_i) \approx \underbrace{\frac{1}{2}\left.\frac{\partial^2 V_i}{\partial z_i^2}\right|_0}_{a_i} z_i^2 + \underbrace{\frac{1}{4!}\left.\frac{\partial^4 V_i}{\partial z_i^4}\right|_0}_{b_i} z_i^4 \tag{2.29}$$

where we have set $V_i(0) = 0$ and eliminated odd-order terms based on crystal symmetry. We also truncate the series at fourth order for simplicity. We emphasize that this energy expansion is not the same as expanding a Landau free energy function. The effective ionic potential energy is quite general here, and its coefficients have no direct correspondence with the order of a ferroelectric transition. The coefficients are calculated from first principles, but we leave their values unspecified for generality. We do assume $a_i, b_i > 0$, however, for ease of discussion.

At this point, we have nearly all of the forces we need to describe our ferroelectric via the dynamics of B cations:

$$\mathbf{F}^\mathrm{B}_\mathrm{ion}(\bar{z}_\mathrm{B}) \approx -(2a_\mathrm{B}\bar{z}_\mathrm{B} + 4b_\mathrm{B}\bar{z}_\mathrm{B}^3)\hat{\mathbf{z}} \tag{2.30}$$

$$\mathbf{F}_p(\mathbf{R}_\mathrm{B}) \approx q_\mathrm{B}\left(-\frac{\mathbf{P}}{\epsilon_0} + \frac{\mathbf{P}}{3\epsilon_0} + \frac{15\mathbf{P}}{4\pi\epsilon_0}\right) \tag{2.31}$$

$\mathbf{F}^\mathrm{B}_\mathrm{ion}(\bar{z}_\mathrm{B})$ is the force on an average B ion due to the local effective interatomic forces where $\bar{z}_\mathrm{B}$ is the mean B ion displacement; $\mathbf{F}_p$ is the force due to the collective dipole field; and q_B is the effective charge of a B ion. If we now consider the short circuit condition, then

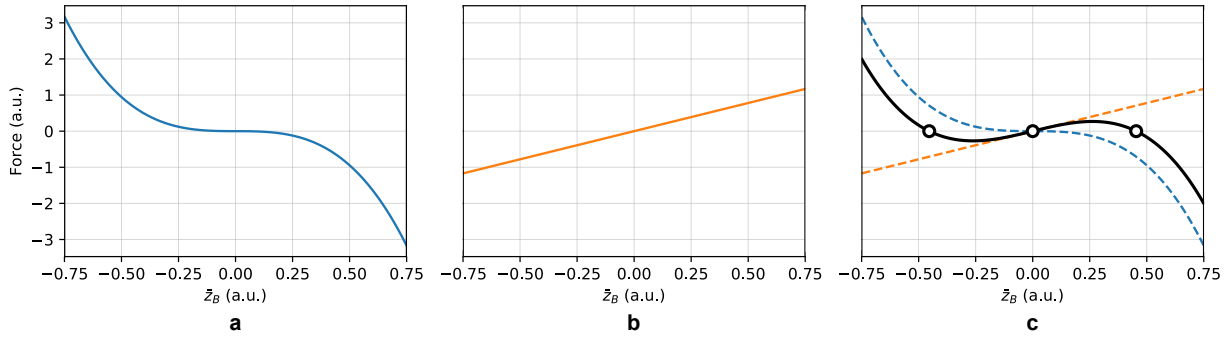

Figure 2.2: Different forces on a B cation. $\bar{z}_{\mathrm{B}}$ is the mean displacement of the B ions. (a) Force due to the local effective ionic potential energy. This force normally stabilizes B ions at zero displacement. (b) Force due to the collective dipole field after the depolarization field has been screened by free charges. This force tends to spontaneously displace B ions, but will be reduced if the depolarization field is not screened. (c) Net force (black) where we have assumed that the local effective interatomic forces are not strong enough to stabilize B ions at zero displacement. The net force is zero at three points but is only stable at two of the points. This resembles the S curve rotated on its side. Without free charges to screen the depolarization field, the dipole field will be weaker, and the interatomic forces may be able to stabilize the B ions. In that case, the ferroelectric will not spontaneously polarize in a vacuum.

free charge will flow and screen the depolarization field. Note that this is a pragmatic approximation since all ferroelectrics have finite leakage in practice and will eventually be in a "short circuit" condition at thermal equilibrium. The macroscopic electric field then becomes $\mathbf{E} = (\mathbf{D} - \mathbf{P})/\epsilon_0 = 0$ where $\mathbf{D}$ is the electric flux density. Thus, there is now an external electric field $\mathbf{D}/\epsilon_0$ that provides an external force

$$\mathbf{F}_{\mathrm{ext}} = q_{\mathrm{B}} \left(\frac{\mathbf{D}}{\epsilon_0} \right) \tag{2.32}$$

If we now plot all of the forces for a B cation, then we obtain the results shown in Fig. 2.2. Fig. 2.2a shows the force due to the effective ionic potential energy, which ordinarily stabilizes B cations at zero displacement. Fig. 2.2b shows the force due to the collective dipole field after the depolarization field has been screened by free charges. This force tends to spontaneously displace B ions. If we assume that the effective interatomic forces are not strong enough to stabilize the B ions, then we obtain the net force shown in Fig. 2.2c. Notice that the net force looks like a rotated S.

In Fig. 2.3a, we replace the short circuit with a voltage source and apply an external electric field in the $-\hat{\mathbf{z}}$ direction. This shifts the S curve vertically such that there is only a single mean ion displacement for which the net force is zero and stable—the ferroelectric is poled in the $-\hat{\mathbf{z}}$ direction. In Fig. 2.3b, we switch the direction of the external electric field,

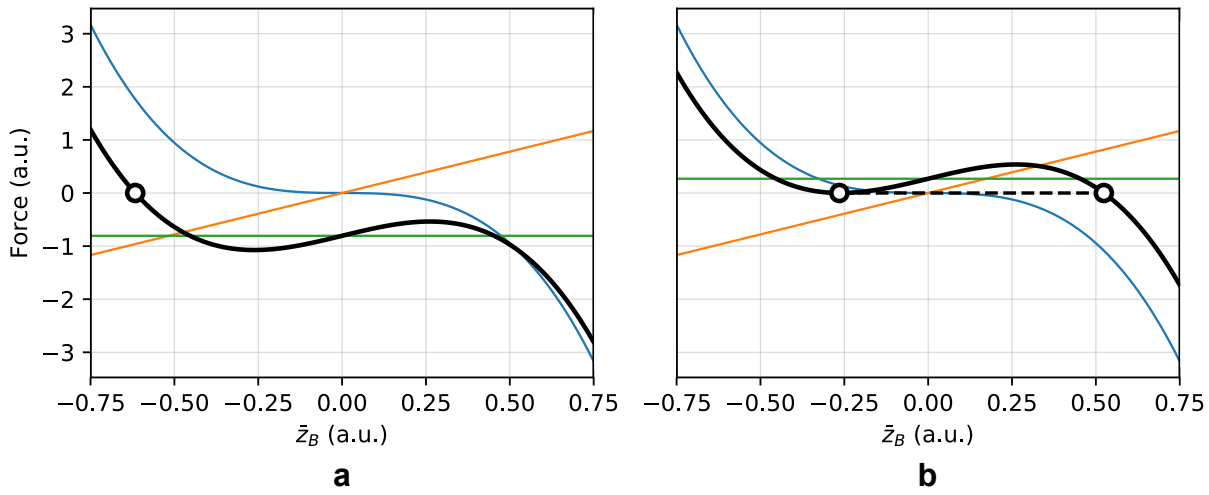

Figure 2.3: Net force (black) on a B cation due to a local effective ionic potential (blue), a collective dipole field (orange), and an external electric field (green). $\bar{z}_{\mathrm{B}}$ is the mean displacement of the B ions. (a) A negative external electric field is applied, forcing the B ions to displace in the $-\hat{\mathbf{z}}$ direction on average. (b) The external electric field is switched and set to the coercive field; the net force becomes zero at two points. Only one point is stable, so the B ions displace in the $+\hat{\mathbf{z}}$ direction on average.

and set its value to the coercive field. Notice that there are now two mean ion displacements for which the net force is zero, but only one is stable. Thus, the ions will displace in the $+\hat{\mathbf{z}}$ direction on average, and the polarization will switch.

Finally, we can express the net force in terms of polarization and macroscopic electric field. The net force is

$$\mathbf{F}^{\mathrm{B}}_{\mathrm{ion}} + \mathbf{F}_p + \mathbf{F}_{\mathrm{ext}} = -(2a_{\mathrm{B}}\bar{z}_{\mathrm{B}} + 4b_{\mathrm{B}}\bar{z}_{\mathrm{B}}^3)\hat{\mathbf{z}} + q_{\mathrm{B}}\left(\frac{\mathbf{P}}{3\epsilon_0} + \frac{15\mathbf{P}}{4\pi\epsilon_0}\right) + q_{\mathrm{B}}\underbrace{\left(\frac{\mathbf{D}-\mathbf{P}}{\epsilon_0}\right)}_{\mathbf{E}} \tag{2.33}$$

As discussed previously, $\langle \mathbf{p}_{\mathrm{B}} \rangle \approx \langle \mathbf{p}_{\mathrm{O}_\parallel} \rangle$, so we can relate the mean B ion displacement with polarization via $\mathbf{P} = 2q_{\mathrm{B}}\bar{z}_{\mathrm{B}}\hat{\mathbf{z}}/v_{\mathrm{cell}}$. If we now set the net force to zero and solve for the macroscopic electric field, then we obtain the following relationship along the z-axis:

$$E = \left(\frac{a_{\mathrm{B}}v_{\mathrm{cell}}}{q_B^2} - \frac{1}{3\epsilon_0} - \frac{15}{4\pi\epsilon_0}\right)P + \left(\frac{b_{\mathrm{B}}v_{\mathrm{cell}}^3}{2q_{\mathrm{B}}^4}\right)P^3 \tag{2.34}$$

This is the S curve (Fig. 2.4a). Thus, we have demonstrated that every point in the S curve corresponds to microscopic configurations of relative ion displacements. If we integrate the electric field with respect to polarization, then we obtain the potential energy landscape shown in Fig. 2.4b. This is a double-well potential and resembles the standard Landau

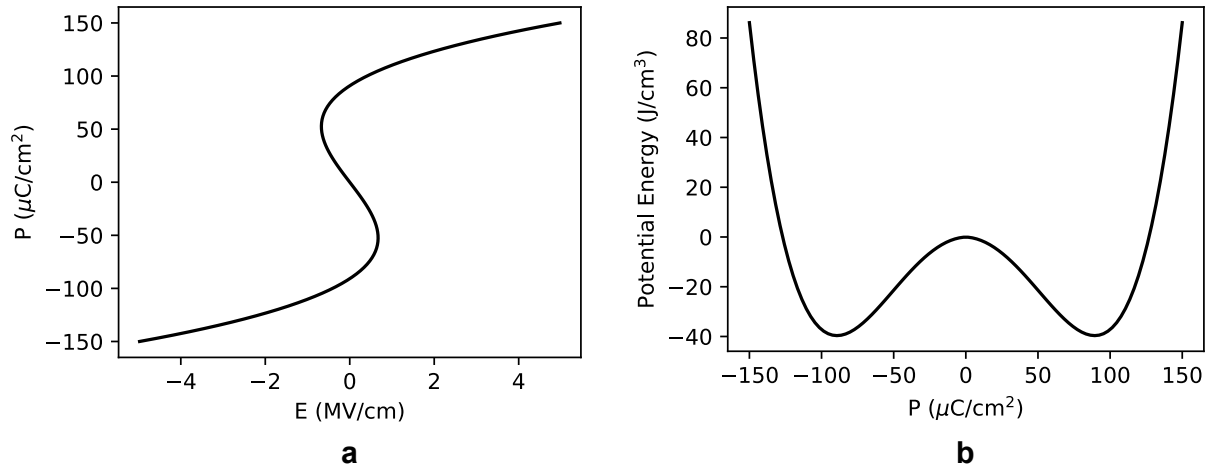

Figure 2.4: (a) Polarization versus electric field corresponding to Fig. 2.3. This is the S curve, and every point on it can be mapped to a microscopic configuration of relative ion displacements. (b) Corresponding potential energy landscape. This double-well potential does not take into account thermal energy and is not a thermodynamic potential, but it resembles the standard Landau free energy landscape.

free energy landscape. Of course, our analysis has assumed zero temperature, and thermal energy will reduce the accessibility of the configurations considered thus far. We can correct for this by taking into account the kinetic energies of the ions and writing down an effective Hamiltonian. The true free energy can then be determined from the canonical partition function, and it will predict a second-order phase transition under the assumptions made in this section (e.g. $a_i, b_i > 0$). We also neglected lattice distortions, effectively removing electromechanical coupling (i.e. piezoelectricity and electrostriction). Adding elastic interactions to the model could result in additional structural transitions (e.g. cubic $\rightarrow$ tetragonal $\rightarrow$ orthorhombic).

2.3 Stabilization in the Negative Capacitance Region

Let us now consider the gate stack of a MOSFET with a ferroelectric gate insulator. For simplicity, let us assume that the semiconductor capacitance is replaced by a linear dielectric capacitor. Then from charge conservation,

$$D_{\mathrm{FE}} = \epsilon_0 E_{\mathrm{FE}} + P = \epsilon_0 \epsilon_{\mathrm{r}} E_{\mathrm{s}} = D_{\mathrm{s}} \tag{2.35}$$

where D_{FE} and D_{s} are the electric flux densities along the gate axis for the ferroelectric and dielectric respectively; E_{FE} and E_{s} are the electric fields across the ferroelectric and dielectric respectively; and ϵ_{r} is the relative permittivity of the dielectric. At $V_{\mathrm{g}} = 0$ (i.e. the short

circuit condition), the dielectric field is related to the ferroelectric field by

$$E_{\mathrm{s}} = -\left(\frac{t_{\mathrm{FE}}}{t_{\mathrm{s}}}\right) E_{\mathrm{FE}} \tag{2.36}$$

where t_{FE} and t_{s} are the thicknesses of the ferroelectric and dielectric respectively. This can be used with (2.35) to obtain the electric field across the ferroelectric:

$$E_{\mathrm{FE}} = \frac{-P/\epsilon_0}{1 + \epsilon_{\mathrm{r}}\left(\frac{t_{\mathrm{FE}}}{t_{\mathrm{s}}}\right)} \tag{2.37}$$

In other words, the series combination leads to a depolarization field on the ferroelectric layer. Since $E_{\mathrm{FE}} = (D_{\mathrm{FE}} - P)/\epsilon_0$, we can also solve for the charge $Q = D_{\mathrm{FE}} = D_{\mathrm{s}}$ on the plates:

$$Q = \frac{P}{1 + \frac{1}{\epsilon_{\mathrm{r}}}\left(\frac{t_{\mathrm{s}}}{t_{\mathrm{FE}}}\right)} \tag{2.38}$$

Notice that the charge decreases when we add a dielectric and further decreases if we increase the dielectric thickness. Thus, the depolarization field arises because we are effectively reducing the screening charge when we add a dielectric. Fig. 2.5a shows the effect of this depolarization field on the force-displacement diagram. Essentially it adds another contribution to the depolarizing (ionic potential) part of the field. As a result, the net force in Fig. 2.2 curves towards zero at smaller deflections or polarizations. In other words, the effect of the depolarization field is the suppression of polarization. The deleterious effect of the depolarization field has long been known in the context of ferroelectric memory devices [61, 62, 63] and standard memory design concepts emphasize on ways to suppress this field. For example from (2.37), we can see that increasing the thickness of the ferroelectric layer with respect to the dielectric layer would be one direct way to suppress the depolarization field. In this context, significant research has gone into unintended dielectric formation (often called the "dead layer") between a ferroelectric and a metal electrode and its effect on suppression of polarization through the generation of a depolarization field. One solution that has been proposed to resolve this issue is the use of an oxide electrode that prevents the formation of a dead layer [64].

Returning to negative capacitance, Fig. 2.5b shows a situation in which the depolarization field is so strong that the only available solution where the net force is zero is also where the polarization is zero. This is where we would like to be for negative capacitance operation. On the unperturbed force-displacement curve, we can see that if we apply an external field to increase the polarization from this point, then the ferroelectric will be acting in the regime in which the net polarizing field is larger than the net depolarizing field inside the ferroelectric. In other words, the material will exhibit enhanced polarization. This enhancement is what boosts the electric field at the ferroelectric-dielectric interface.

To complete this discussion, we show another situation in Fig. 2.5c in which the depolarization field suppresses the polarization close to zero but not exactly to zero. Notice that

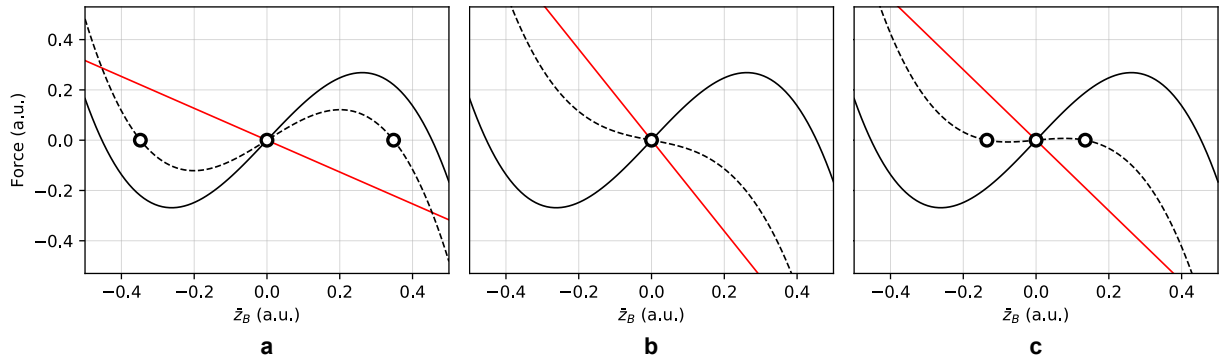

Figure 2.5: Net force on a B cation when a depolarization field is present. The solid black curve is the force without the depolarization field; the red line is the force due to the depolarization field; and the dashed black curve is the net force. (a) Depolarization field slightly reduces the remnant polarization. (b) Depolarization field is strong enough to suppress the polarization to zero. (c) Depolarization field suppresses polarization close to zero but not quite zero. In this case, the combination is still in the negative capacitance region but will have hysteresis.

there are two stable points again where the net force can be zero. In other words, there will be hysteresis. However, the polarization is still suppressed enough that the ferroelectric is stabilized in a region where the net polarizing field wins over the net depolarizing field. Thus, the ferroelectric exhibits negative capacitance albeit with small hysteresis. This region may be unattractive for logic applications but is, nonetheless, important to understand and analyze in experimental data.

Let us now examine how strong a depolarization field is needed to stabilize the ferroelectric at zero polarization as in Fig. 2.5a. We have already derived a relationship between electric field and polarization for a ferroelectric in (2.34), but the addition of a dielectric imposes the constraint derived in (2.37):

$$E_{\mathrm{FE}} = \left(\frac{a_{\mathrm{B}} v_{\mathrm{cell}}}{q_B^2} - \frac{1}{3\epsilon_0} - \frac{15}{4\pi\epsilon_0}\right) P + \left(\frac{b_{\mathrm{B}} v_{\mathrm{cell}}^3}{2q_{\mathrm{B}}^4}\right) P^3 = \frac{-P/\epsilon_0}{1 + \epsilon_{\mathrm{r}}\left(\frac{t_{\mathrm{FE}}}{t_{\mathrm{s}}}\right)} \tag{2.39}$$

We can solve this equation for the possible polarizations, and there are three solutions in general:

$$P = \begin{cases} 0 \\ \pm\sqrt{\frac{2q_{\mathrm{B}}^4}{b_{\mathrm{B}} v_{\mathrm{cell}}^3}}\sqrt{\left(\frac{1}{3\epsilon_0} + \frac{15}{4\pi\epsilon_0} - \frac{a_{\mathrm{B}} v_{\mathrm{cell}}}{q_{\mathrm{B}}^2}\right) - \frac{1/\epsilon_0}{1+\epsilon_{\mathrm{r}}\left(\frac{t_{\mathrm{FE}}}{t_{\mathrm{s}}}\right)}} \end{cases} \tag{2.40}$$

Stabilizing the ferroelectric at zero polarization requires us to eliminate the two non-zero solutions. We can accomplish this by making them imaginary, which occurs under the

condition

$$\frac{1/\epsilon_0}{1+\epsilon_r\left(\frac{t_{FE}}{t_s}\right)} > \frac{1}{3\epsilon_0}+\frac{15}{4\pi\epsilon_0}-\frac{a_B v_{cell}}{q_B^2}$$

or equivalently,

$$\frac{\epsilon_0\epsilon_r}{t_s} < \frac{1}{t_{FE}}\left(\frac{1}{\frac{1}{3\epsilon_0}+\frac{15}{4\pi\epsilon_0}-\frac{a_B v_{cell}}{q_B^2}}-\epsilon_0\right) \tag{2.41}$$

Note that this condition is equivalent to the stability condition originally implied in (2.13).

The fact that stabilization in the negative capacitance region happens for values of P where the potential energy versus polarization curve (Fig. 2.4b) has a negative curvature has often led to some confusion in the community as to how one can obtain a stable state in this seemingly "unstable" region. *The point of the above analysis is to show that the ferroelectric becomes stable in a region of the configuration space in which the net force on an average dipole is zero. These configurations, in fact, correspond to the most stable overall state for the ferroelectric-dielectric series combination although they might correspond to an unstable state for an unperturbed (isolated) ferroelectric. The second key point that we want to emphasize is the fact that every state in this negative capacitance regime is associated with physically defined structures of the unit cell.*

The above analysis also provides some important insights:

1. It is more advantageous for negative capacitance operation if the ferroelectric can smoothly traverse the configuration space to transition to a zero polarization state. In that context, a second order phase transition material seems more appropriate than a first order phase transition material.

2. In the same token, if a substrate strain is present, it may be difficult for the ferroelectric to transition to zero polarization as it will involve acting against the substrate strain. Therefore, for epitaxially grown heterostructures, it may be more preferable for the ferroelectric to break into domains and stabilize the negative capacitance state in the domain walls. In fact, this has recently been observed in epitaxially grown superlattices. On the other hand, such constraints from the substrate are not present for polycrystalline ferroelectrics, and, therefore, it might be much easier for the entire volume of the material to stabilize into the negative capacitance state. We therefore believe that polycrystalline ferroelectrics such as doped HfO_2 are more appropriate for negative capacitance operation. We suggest that one should sandwich such a ferroelectric layer with an amorphous dielectric layer, which should more easily accommodate changes in strain.

2.4 Breaking the Fundamental Energy Dissipation Limit

Half of the energy is always lost when charging a capacitor [65]. Even in the limit of vanishing resistance, half of the charging energy is still lost—to radiation instead of heat. While this fraction can technically be reduced by charging adiabatically, it otherwise places a fundamental limit on the charging efficiency of a capacitor. Here we show that this 1/2 limit can be broken by coupling a ferroelectric to the capacitor dielectric. Maxwell's equations are solved for the coupled system to analyze energy flow from the perspective of Poynting's theorem and show that (1) total energy dissipation is reduced below the fundamental limit during charging and discharging; (2) energy is saved by "recycling" the energy already stored in the ferroelectric phase transition; and (3) this phase transition energy is directly transferred between the ferroelectric and dielectric during charging and discharging. These results demystify recent works [66, 20, 67, 19, 21, 68, 22, 23] on low energy negative capacitance devices as well as lay the foundation for improving fundamental energy efficiency in all devices that rely on energy storage in electric fields.

Ferroelectrics have recently been shown to exhibit a negative capacitance effect [18, 20, 23] when placed in a series combination with a dielectric film. Under appropriate conditions, the dielectric leads to a strong depolarization field that forces the ferroelectric into its normally unstable near-zero polarization states. These results have garnered considerable interest in ferroelectric negative capacitance due to its potential to reduce power consumption below thermodynamic limits in electronic devices [18]. In field-effect transistors, for example, negative capacitance has been proposed as a solution to end the "Boltzmann tyranny" on subthreshold swing [18, 43, 44]. However, negative capacitance has traditionally been understood to require a source (e.g. a battery) that supplies extra energy [69]. This begs the following question: if extra energy must be supplied by some source, then where is the energy coming from in ferroelectric negative capacitance and is any energy truly saved? To answer this question, we considered energy flow during charging and discharging of a ferroelectric-dielectric capacitor. We solved Maxwell's equations for the coupled system and used Poynting's theorem to show how energy flows.

2.4.1 Poynting's Theorem

Poynting's theorem is a statement of conservation of energy for a system of charged particles and can be written as a differential continuity equation:

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{S} = -\mathbf{J} \cdot \mathbf{E} \tag{2.42}$$

The integral form is

$$\int_\Omega \left(\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{S} \right) d^3r = -\int_\Omega (\mathbf{J} \cdot \mathbf{E})\, d^3r \tag{2.43}$$

where the volume Ω can be arbitrary, but we take it as the volume occupied by the components of the circuit to establish appropriate boundary conditions. This differential equation can be numerically solved for the current density $\mathbf{J}$ if the remaining variables can be expressed in terms of current density. We accomplish this by establishing a consistent relationship between internal energy density u and electric field $\mathbf{E}$ for a given current density. First, an initial electric field is assumed, and the state of the ferroelectric-dielectric system is determined by minimizing the free energy in (2.49) with respect to small polarization fluctuations. This determines the ferroelectric and dielectric polarization states, which allow us to determine the internal energy density u and charge density ρ. This charge density can be used in conjunction with a given current density to solve Maxwell's equations by determining the retarded scalar and vector potentials:

$$V(\mathbf{r},t) = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\mathbf{r}',t')}{|\mathbf{r}-\mathbf{r}'|}\delta\left(t'-t+\frac{|\mathbf{r}-\mathbf{r}'|}{c}\right)\,dt'\,d^3r' \tag{2.44}$$

$$\mathbf{A}(\mathbf{r},t) = \frac{\mu_0}{4\pi}\int \frac{\mathbf{J}(\mathbf{r}',t')}{|\mathbf{r}-\mathbf{r}'|}\delta\left(t'-t+\frac{|\mathbf{r}-\mathbf{r}'|}{c}\right)\,dt'\,d^3r' \tag{2.45}$$

The solutions to Maxwell's equations are then

$$\mathbf{E}(\mathbf{r},t) = -\nabla V - \frac{\partial \mathbf{A}}{\partial t} \tag{2.46}$$

$$\mathbf{B}(\mathbf{r},t) = \nabla \times \mathbf{A} \tag{2.47}$$

To reduce numerical error, we directly solved Maxwell's equations by analytically combining (2.44)-(2.47) and evaluating the resulting Jefimenko's equations. Notice that the electric field computed here must be equal to the electric field initially assumed when determining the internal energy density. Thus, the electric field and internal energy density must be solved for consistently, and this can be accomplished for a given current density. From (2.45) and (2.47), the magnetic field $\mathbf{B}$ only depends on current density, so the Poynting vector $\mathbf{S}$ can also be computed for a given current density:

$$\mathbf{S} = \frac{\mathbf{E}\times\mathbf{B}}{\mu_0} \tag{2.48}$$

The only remaining unknown quantity is the current density, which can now be solved for using Poynting's theorem.

2.4.2 Energy Flow and Dissipation

We performed our analysis in the limit of zero resistance in order to understand the fundamental charging and discharging energy costs. The results are shown schematically in Fig. 2.6a. During charging, input energy flows from an energy source to the dielectric, and a fraction of that energy is dissipated. This dissipation is dominated by electromagnetic radiation in the limit of zero resistance. Notice that there is an additional path of energy flow

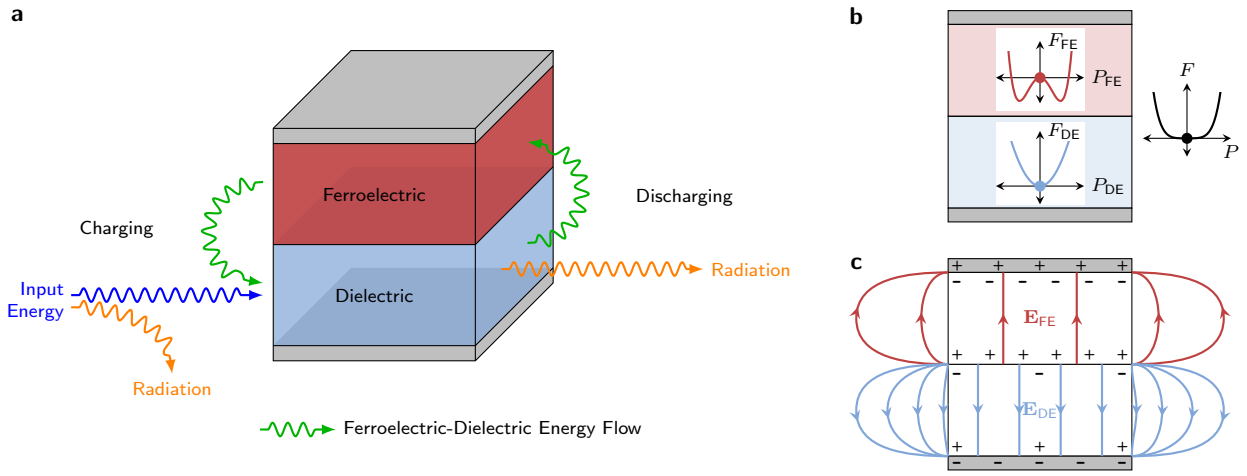

Figure 2.6: **Paths of energy flow during charging and discharging of a capacitor with a ferroelectric.** **a**, Schematic of energy flow paths during charging (left) and discharging (right) of a ferroelectric-dielectric capacitor. New paths of energy flow emerge between the ferroelectric and dielectric during charging and discharging. These paths are not present in conventional dielectric capacitors. **b**, Energy landscapes in a ferroelectric-dielectric capacitor. If the dielectric is sufficiently thick, then its energy landscape F_{DE} will dominate the energy landscapes F_{FE} and F of the ferroelectric and overall system respectively. The dielectric polarization P_{DE} then forces the ferroelectric near its phase transition at zero polarization ($P_{\mathrm{FE}} = 0$) via a strong depolarization field. This puts the ferroelectric into a higher energy state in which energy can be extracted from the phase transition. **c**, Schematic of the total electric fields in a ferroelectric-dielectric capacitor. When an external electric field is applied, the ferroelectric and dielectric both polarize by different amounts, resulting in a depolarization field. Since the ferroelectric is stabilized in a higher energy state near zero polarization, it releases energy when polarized. This extra energy contributes towards further strengthening the depolarization field, which subsequently further polarizes and charges the dielectric. The resultant electric fields $\mathbf{E}_{\mathrm{FE}}$ and $\mathbf{E}_{\mathrm{DE}}$ in the ferroelectric and dielectric respectively end up pointing in opposite directions.

from the ferroelectric to the dielectric that is not present during charging in conventional capacitors. This implies that the ferroelectric is supplying extra energy to the dielectric. During discharging, the dielectric acts as the energy source and normally dissipates all of its energy when in a conventional capacitor. However, there is a new path of energy flow that allows the dielectric to transfer a fraction of its energy back to the ferroelectric. Thus, we see schematically how energy may be internally recycled in the coupled ferroelectric-dielectric system. However, it is still unclear where the extra energy comes from and how it transfers between the ferroelectric and dielectric.

The origin of this extra energy can be understood from a thermodynamic perspective as shown in Fig. 2.6b. Due to their phase transition, ferroelectrics possess a higher energy, zero polarization state in their energy landscape. This is in contrast to dielectrics, which have a minimum in their energy landscape at zero polarization. Consequently, coupling a ferroelectric to a dielectric results in a large divergence in polarization at the interface. This establishes a strong depolarization field that stabilizes the ferroelectric near its unstable zero polarization state. The net effect is an electrically-induced transition towards a phase of higher crystal symmetry and can be thought of as an effective shift in the phase transition temperature [18, 23]. This electrical influence is in conflict with the natural temperature-induced transition towards lower crystal symmetry. Thus, we can electrically extract energy from the phase transition by modulating this conflict with an applied electric field. The extracted energy is then transferred between the ferroelectric and dielectric via propagation of the depolarization field as shown schematically in Fig. 2.6c. Notice that the electric field points in opposite directions from the ferroelectric-dielectric interface due to the negative permittivity of the ferroelectric near its phase transition. It is worth noting that this result was directly obtained from our calculations without any consideration a priori of negative electric susceptibilities or capacitances.

For our quantitative analysis, we modelled the coupled ferroelectric-dielectric system using the electric Gibbs free energy (which we will refer to as simply free energy for the remainder of this section):

$$G = \int (f - \mathbf{E} \cdot \mathbf{P}) \, d^3r \tag{2.49}$$

f is the Helmholtz free energy density as a function of temperature and polarization $\mathbf{P}$; and $\mathbf{E}$ is electric field. Note that f, $\mathbf{E}$, and $\mathbf{P}$ all vary with position $\mathbf{r}$, and the functional form of f depends on the material energy landscape. The dielectric was modelled as a linear dielectric, and its electric susceptibility and thickness were normalized as a single tuning parameter. The ferroelectric energy landscape was modelled after $Pb(Zr_{0.52}Ti_{0.48})O_3$ using Landau-Devonshire phenemonological parameters [57]. We could have used Ginzburg-Landau theory to take into account slow spatial variations in the polarization. However, such fine details would simply add finer spatial variations to our calculated energy flow; the overall flow would remain the same as long as the ferroelectric was still locally stabilized near its zero polarization state. We also assumed a one-dimensional order parameter since polarization is expected to lie primarily along the capacitor axis. Finally, we solved for

the stationary states of the coupled ferroelectric-dielectric system by minimizing the free energy with respect to small polarization fluctuations $\delta\mathbf{P}$ under constant electric field and isothermal conditions.

The dynamics of the system were described with Poynting's theorem (2.42). The internal energy density u inside the ferroelectric and dielectric was determined by solving for the states of the coupled ferroelectric-dielectric system (as described in the previous paragraph). The form of the Poynting vector $\mathbf{S}$ was obtained by solving Maxwell's equations using retarded scalar and vector potentials. For ease of calculation, we considered a simple wire loop geometry containing a voltage source and the ferroelectric-dielectric capacitor at diametrically opposite positions. With u and $\mathbf{S}$, it is possible to numerically solve the differential equation (2.42) and compute the power supplied and consumed over time as shown in Fig. 2.7a-b. During charging (Fig. 2.7a), less power is supplied by the energy source when a ferroelectric is coupled to the dielectric, and less power is lost to radiation. However, the dielectric still receives the same amount of energy, and we find that the ferroelectric supplies the missing part using the energy stored in its phase transition. During discharging (Fig. 2.7b), the dielectric radiates energy, but the ferroelectric appears to "capture" a fraction of it back to replenish its phase transition energy when coupled to the dielectric. The total energy dissipated is shown in Fig. 2.7c as a function of the "capacitance matching" between the ferroelectric and dielectric. We find that the amount of energy dissipated is reduced below the conventional $^1/_2\,CV^2$ during charging and discharging, and it is minimized when the ferroelectric and dielectric capacitances are equal (i.e. $|C_{\mathrm{FE}}/C_{\mathrm{DE}}| \to 1$). It should be noted that the capacitor becomes nonlinear when a ferroelectric is added (inset of Fig. 2.7c), resulting in charge-dependent energy dissipation. We charged to the end of the linear region and then discharged completely to calculate the curves in Fig. 2.7c.

The capacitance matching can be better understood by examining the energy landscapes shown in Fig. 2.8a. In the left set of landscapes, the energy available in the ferroelectric phase transition closely matches the energy "needs" of the dielectric in the linear region. This allows the ferroelectric to supply nearly all of the energy needed to charge the dielectric. Consequently, minimal additional energy is needed from an external source, and less energy will be lost to radiation while propagating from the source to the dielectric. This is an example of perfect energy balancing. In contrast, the right half of Fig. 2.8a shows an example of poor energy balancing. The energy available in the ferroelectric is insufficient for charging the dielectric. Consequently, more energy is needed from an external source, and more energy will be lost to radiation while propagating to the dielectric. We can control the energy balancing by tuning the energy landscapes. This is accomplished by changing film thickness or electric susceptibility (e.g. by changing temperature or strain; or using different materials). Since these parameters directly affect the system's capacity to store energy in an electric field, the energy balancing can be thought of as a capacitance matching between the ferroelectric and dielectric. In the linear region, for example, the ferroelectric capacitance is negative due to the negative curvature of the energy landscape and should be equal and opposite to the dielectric capacitance for proper matching. The ferroelectric is able to supply energy to the dielectric within this linear region. However, it will eventually

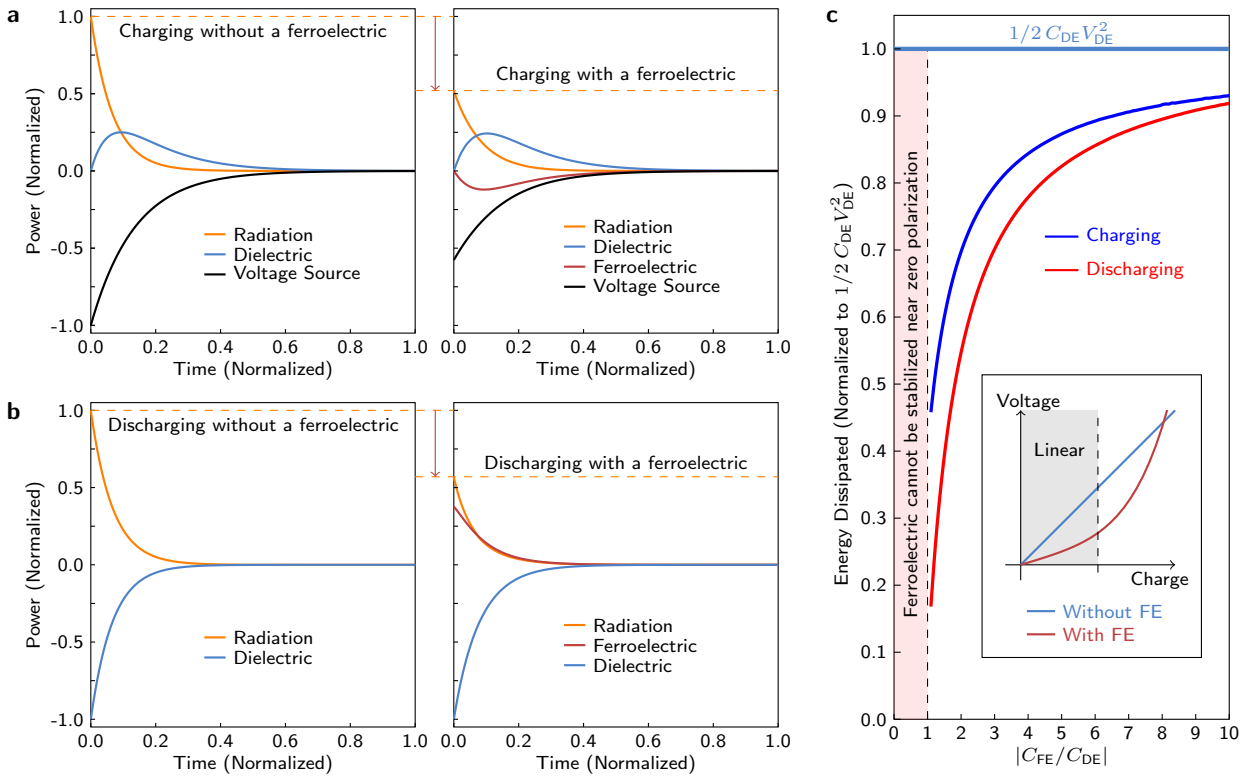

Figure 2.7: **Power comparison during charging and discharging of a capacitor without/with a ferroelectric.** Negative power corresponds to power supplied. **a**, Power versus time during charging of a capacitor without a ferroelectric (left) and with a ferroelectric added (right). The ferroelectric and dielectric parameters are set such that $|C_{FE}/C_{DE}| = 2$ (see **c** and Fig. 2.8 for more information) where C_{FE} is the ferroelectric capacitance, and C_{DE} is the dielectric capacitance. The voltage source supplies less power when the ferroelectric is coupled to the dielectric, and the amount of power radiated is reduced. The dielectric still receives the same amount of energy because the ferroelectric supplies the missing power. **b**, Power versus time during discharging of the same capacitors from **a** (without a ferroelectric, left; and with a ferroelectric, right). The dielectric acts as the source during discharging, and a fraction of its power is delivered to the ferroelectric instead of completely radiating away as in the conventional case. **c**, Total energy dissipated as a function of the capacitance matching $|C_{FE}/C_{DE}|$ after charging and discharging. The energy is normalized to $^1/_2\,C_{DE}V_{DE}^2$, which is the conventional amount of energy dissipated without a ferroelectric. The inset shows that the capacitor becomes nonlinear when a ferroelectric is added, resulting in charge-dependent energy dissipation. The curves shown here were calculated by charging to and discharging from the end of the linear region.

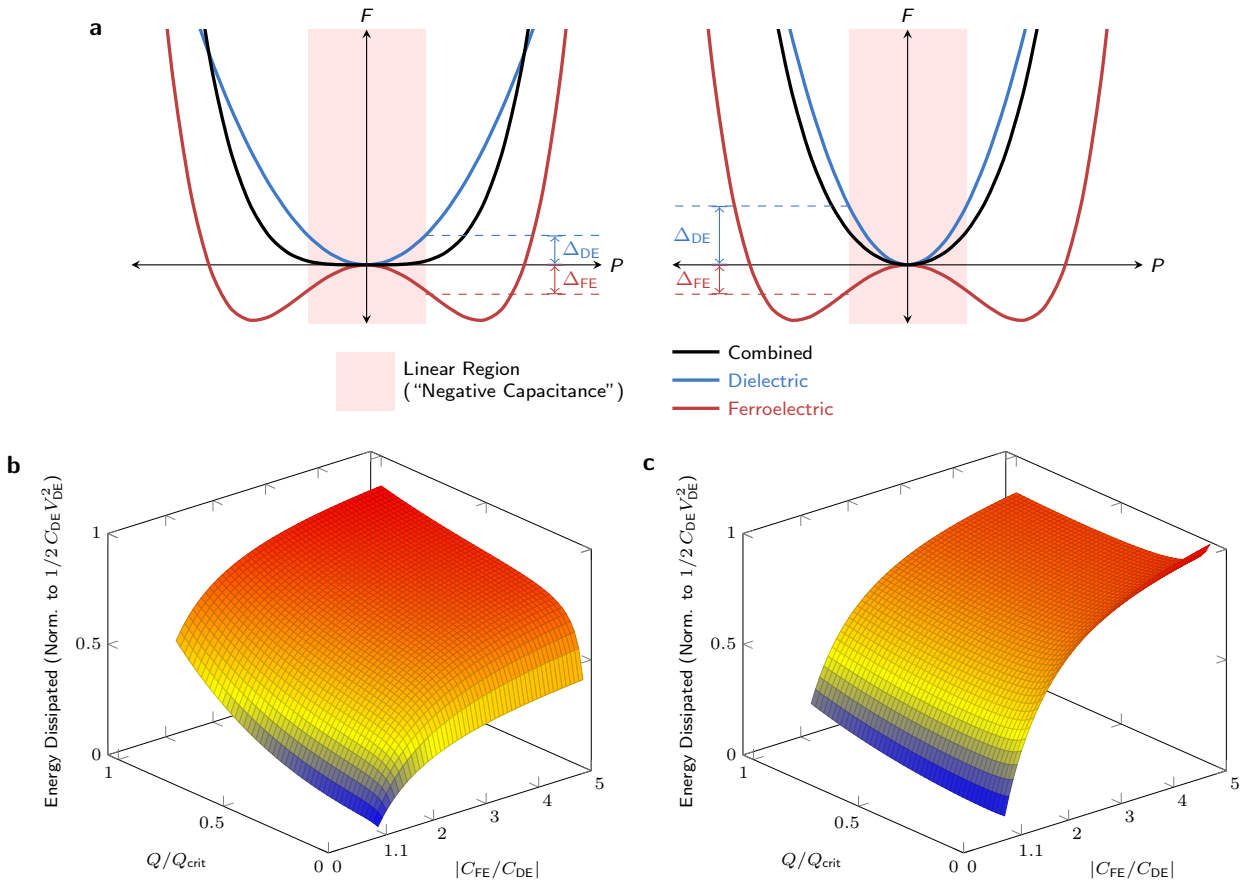

Figure 2.8: **Energy balancing and capacitance matching. a**, Energy landscapes showing perfect energy balancing $\Delta_{DE} \approx \Delta_{FE}$ (left) and imperfect energy balancing $\Delta_{DE} > \Delta_{FE}$ (right) between the ferroelectric and dielectric. Even with perfect energy balancing, the ferroelectric eventually runs out of energy stored in its phase transition. This occurs at the end of the linear region, after which point the ferroelectric can no longer supply energy to the dielectric and must receive energy from an external source to continue polarizing. **b**, Total energy dissipated as a function of the capacitance matching $|C_{FE}/C_{DE}|$ after storing charge Q on the capacitor plates. C_{FE} is the ferroelectric capacitance, and C_{DE} is the dielectric capacitance. The energy dissipated is normalized to $1/2\,C_{DE}V_{DE}^2$, which is the energy conventionally dissipated without a ferroelectric. Q_{crit} is the charge corresponding to the end of the linear region. **c**, Total energy dissipated as a function of the capacitance matching $|C_{FE}/C_{DE}|$ after discharging Q amount of charge.

run out of stored energy and begin requiring energy from an external source to continue polarizing. This is reflected by the change in the energy landscape's curvature from negative to positive at the end of the linear region. This is also reflected in the charge dependency of the energy dissipation as shown in Fig. 2.8b-c. No matter how perfectly matched the ferroelectric and dielectric are, the energy dissipation increases for greater amounts of charge during both charging (Fig. 2.8b) and discharging (Fig. 2.8c).

Finally, we explicitly show how energy transfers between the ferroelectric and dielectric by computing the Poynting vector at various positions in and around the system. The overall energy flow during charging is shown schematically in Fig. 2.9a. Notice that energy flows from the source to the ferroelectric-dielectric system, but some of it radiates away. If we zoom in on the capacitor (Fig. 2.9b), we see that energy flows directly from the ferroelectric to the dielectric through the surrounding space. The Poynting vector diverges outwards (energy is decreasing) at the center of the ferroelectric and inwards (energy is increasing) at the dielectric center. During discharging (Fig. 2.9c), the Poynting vector diverges oppositely compared to the charging case. The dielectric acts as the source with some of its energy flowing into the ferroelectric through the surrounding space while the remainder radiates away.

In conclusion, we have shown that it is possible to improve upon the otherwise fundamental limit on energy dissipation of $^1/_2\,CV^2$ during charging and discharging of a capacitor by coupling a ferroelectric to the dielectric. We used a thermodynamic model to show that the dielectric can stabilize the ferroelectric near its phase transition, enabling extraction of the energy stored in the phase transition. Poynting's theorem and Maxwell's equations then explicitly showed that this stored energy directly flows between the ferroelectric and dielectric during charging and discharging. The net result is a reduction in total energy dissipation below the conventional limit. This reduction can be maximized by balancing the energy stored in the ferroelectric phase transition with the energy needed by the dielectric. These results provide the framework for understanding and improving fundamental energy efficiency in all devices that operate on storing energy in electric fields.

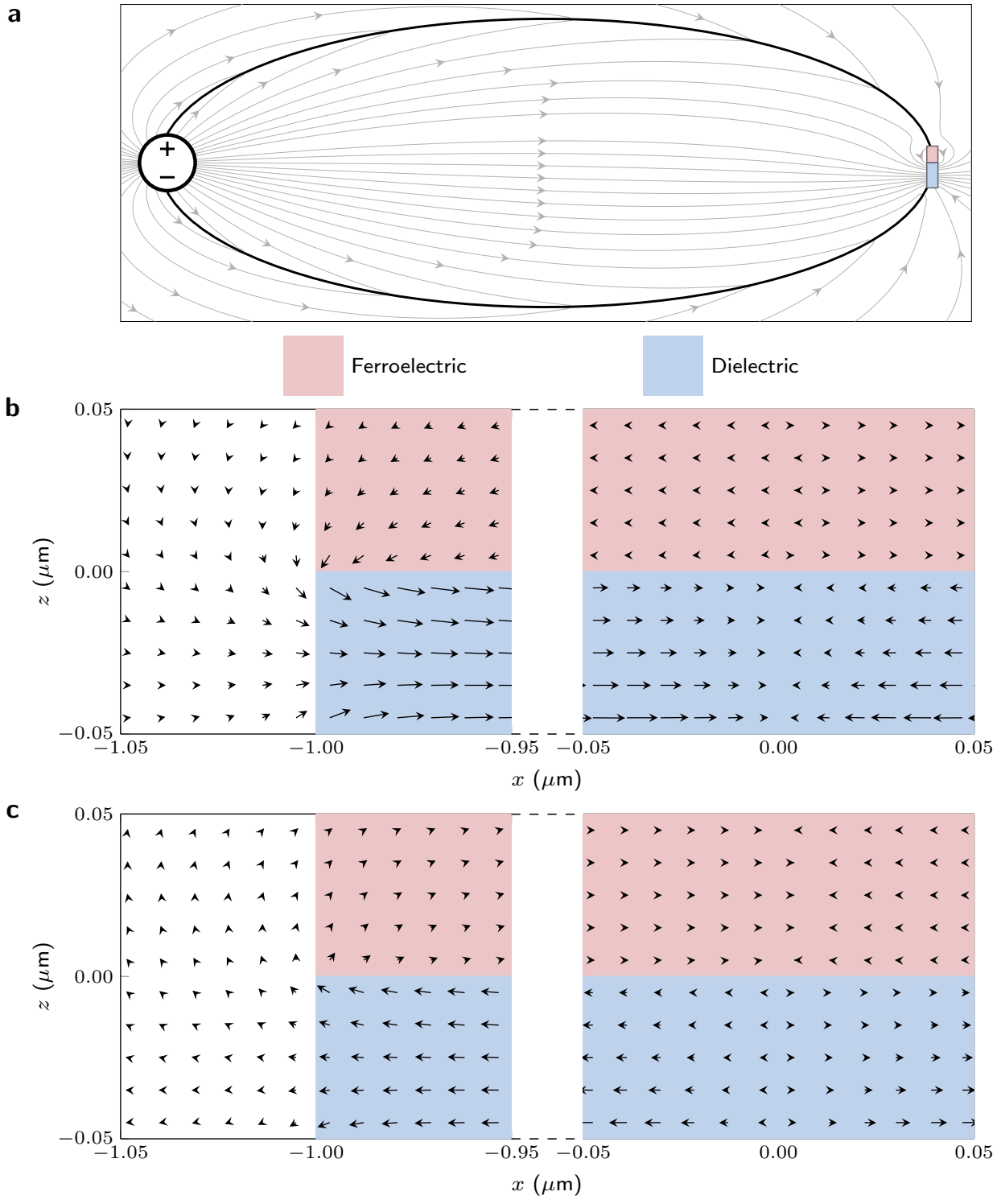

Figure 2.9: **Poynting vector calculations.** **a**, Schematic of overall energy flow from the source to the ferroelectric-dielectric capacitor during charging. **b**, Poynting vector field along the side of the capacitor and near the center during charging. x is the in-plane spatial coordinate, and z is the spatial coordinate along the capacitor axis. **c**, Poynting vector field along the side of the capacitor and near the center during discharging.

Chapter 3

It's Too Noisy

3.1 Introduction

In Chapter 2, we discussed how negative capacitance can, in principle, be used to reduce the body factor in metal-oxide-semiconductor field-effect transistors (MOSFETs) below the presumed ideal limit. Instead of striving for perfect capacitive coupling between the gate and channel, negative capacitance goes beyond by effectively amplifying the gate voltage at the channel. This circumvents the Boltzmann limit on subthreshold swing, allowing for reduced threshold voltage without increasing leakage current or sacrificing I_{on}/I_{off} (the ratio of transistor ON state current to OFF state current). Thus, negative capacitance helps soften the subthreshold barrier on supply voltage scaling. Unfortunately, however, this is still insufficient for enabling arbitrary reductions in supply voltage.

There are other nonscaling barriers that prevent supply voltage scaling. For example, temperature is still nonscaling and, consequently, imposes other thermal limits (aside from the Boltzmann limit on subthreshold swing) such as thermal noise. Even with the regenerative properties of complementary metal-oxide-semiconductor (CMOS) technology, inverters and NAND/NOR gates still require supply voltages at least $4 \sim 10$ times (depending on fan-in) the thermal voltage to maintain logical consistency [24, 25, 17]. Negative capacitance can help relax this constraint by improving voltage transfer characteristics to increase noise margins, but the noise margins will still be limited by voltage swing even with infinite gain and perfect switching. These noise margins must still be large enough to overcome other nonscaling internal noise sources that cannot be easily rejected with clever circuit designs. These requirements are even stricter in memory arrays, which are in lower signal-to-noise ratio (SNR) situations due to additional noise sources such as crosstalk [25, 26]. Furthermore, process tolerance results in variation in device properties such as channel length, oxide thickness, and threshold voltage. This variation worsens with scale unless process tolerances are improved proportionally. Thus, the threshold voltage and supply voltage must be kept sufficiently high to maintain necessary noise margins and SNR.

In this chapter, we use the concept of high-dimensional distributed representations to

alleviate the noise problems mentioned above. We first discuss the implications of choosing a particular representation for information. Then, we highlight the differences between local and distributed representations. Local representations can have certain advantages but often result in computational intractability due to explosions in representational complexity. In contrast, distributed representations can be used to perform computations with more efficient space complexity and are robust against errors. Increasing the dimensionality of distributed representations further increases their robustness against errors.

3.2 Representations of Information

Representations of information—or *knowledge representations*—are so ubiquitous that we often fail to appreciate their importance or consider their consequences. As an obvious example, in computer science we often choose different data structures to represent different kinds of information. This choice of data structure strongly influences the rules that must then be followed to subsequently transform the represented information. Queues and stacks, for example, push and pop elements in completely different fashions. Using one over the other can make certain problems more readily solvable while making others more difficult. Thus, the choice of representation can significantly affect the kind of information processing that can be performed efficiently. Although we used a computing example here, this same idea extends well beyond the realm of computing.

The familiar *Hindu-Arabic numeral system*, for example, is just one possible representation of the concept of numbers. It consists of an *alphabet*

$$\Sigma = \{0, 1, 2, \ldots, 9\} \tag{3.1}$$

where the elements of this alphabet are the *symbols* of the representation and are referred to as digits. We can then string together these symbols to form different *words* or numbers. As a trivial example,

$$\Sigma^1 = \{0, 1, 2, \ldots, 9\} \tag{3.2}$$

is a set of representations for single digit numbers (not to be confused with Σ, which is the alphabet of digits and not a set of number representations). Similarly,

$$\Sigma^2 = \{00, 01, 02, \ldots, 09, 10, 11, 12, \ldots, 99\} \tag{3.3}$$

is a set of representations for numbers up to two digits. Notice that 0 and 00 are two different representations of the same number zero. In fact, there are infinitely many representations that can be obtained for each number in the Hindu-Arabic numeral system by simply prepending leading zeros. Thus, with

$$\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cdots \tag{3.4}$$

we obtain the set of all representations of all natural numbers $\mathbb{N}$. In other words, we are able to represent every element of an infinite set of numbers using a finite alphabet. To represent

other number systems, we can add more symbols to our alphabet. For example, we can add a negative sign and a decimal point to our alphabet to extend the Hindu-Arabic numeral system to the *decimal numeral system* to represent all real numbers $\mathbb{R}$.

If we now attempt to perform arithmetic, then we find that our choice of number representation has a significant impact on our arithmetic algorithms. For example, the addition of two numbers requires us to map any two numbers to their sum:

$$\mathbb{N} \times \mathbb{N} \to \mathbb{N} \tag{3.5}$$

This map contains an infinite number of elements and would be impossible to memorize. However, by using the finite alphabet in (3.2) to represent all possible natural numbers $\mathbb{N}$, we can construct a simpler map that only adds single digit numbers:

$$\{0, 1, 2, \ldots, 9\} \times \{0, 1, 2, \ldots, 9\} \to \{0, 1, 2, \ldots, 18\} \tag{3.6}$$

Thus, we only need to memorize a map containing 10^2 elements (or half of that if we understand commutativity). To add numbers with arbitrary digits, we simply use the concept of a *carry* digit. Similarly for multiplication, we can memorize a reduced map of products of single digit numbers (i.e. a multiplication table) and then use *long multiplication* to multiply numbers with arbitrary digits.

This concept of alphabets, symbols, and words is the same notion used in the theory of computation where problems are formulated as *languages* [70]. The same problem can be formulated as different languages over different alphabets, but the grammar for transforming words will necessarily change. Thus, by choosing appropriate representations, certain problems can become more readily solvable. As a pivotal example, Claude Shannon's recognition of the logical equivalence of boolean algebra with certain types of electrical circuits [71] helped launch binary representations as the representation of choice in digital computing.

3.2.1 Local Representations

Continuing from the numeral system example used above, we can consider an extreme example of a representation using a countably infinite set of symbols for the alphabet Σ. Thus, by definition, there exists a bijection (i.e. a one-to-one correspondence) between this alphabet and the set of natural numbers $\mathbb{N}$. In other words, we represent each natural number using a single symbol rather than a word composed of multiple symbols. This is an example of a *local representation* because each symbol of the alphabet only codes for a single entity or concept. This "local" notion will hopefully make more sense when we contrast with distributed representations in the next subsection.

Perhaps a more accessible example might be the representation of words in the Chinese writing system. In the Chinese "alphabet", there are at least 85,000 symbols or Chinese characters [72] (though only a fraction are commonly used in practice [73, 74]). This is a local representation because each of the symbols represents a single word. (Technically, modern Chinese characters actually approximately represent syllables [75], and most modern Chinese

words are actually written using two symbols [76]. However, we ignore this for illustrative purposes since this is still on the "local representation" end of the representation spectrum). The primary advantage of this representation is that words and pronunciation are decoupled [77] since words are local to individual symbols. This enables the many different Chinese dialects to all use the same Chinese alphabet as their own written language.

On the other hand, there are many disadvantages with using this kind of representation. The most obvious disadvantage is the need to memorize a large set of symbols (i.e. one symbol per word). This is in analogy with the numeral system example presented at the beginning of this subsection where we used a countably infinite alphabet set to represent the set of natural numbers. Performing arithmetic with such a local representation amounts to finding maps

$$+ : \mathbb{N} \times \mathbb{N} \to \mathbb{N} \tag{3.7}$$

$$\cdot : \mathbb{N} \times \mathbb{N} \to \mathbb{N} \tag{3.8}$$

for the addition $+$ and multiplication $\cdot$ binary operations. As discussed prior to this subsection, these maps contain infinite elements and would be impossible to implement in a computer due to bounded memory. In fact, it is impossible to construct a finite automata over an infinite alphabet.

Another key disadvantage of local representations is their susceptibility to errors. If we incur a single symbol error in a local representation, then the entire representation can change completely. In the Chinese example, a single symbol error results in an entirely different word. Clearly, this type of representation will not be helpful for addressing the noise problems presented at the beginning of this chapter.

3.2.2 Distributed Representations

Let us now contrast the Chinese writing system with the English writing system. In the English writing system, the alphabet contains 26 symbols called letters:

$$\Sigma = \{a, b, \ldots, z\} \tag{3.9}$$

We string together these letters to form words:

$$\Sigma^* = \{a, b, \ldots, z, aa, ab, \ldots\} \tag{3.10}$$

The English language L is then defined as the subset of Σ^* that contains all of the valid English words:

$$L = \{w \in \Sigma^* | w \text{ is a valid English word}\} \tag{3.11}$$

Thus, each word is represented using multiple symbols. This is an example of a *distributed representation*. In a distributed or "holographic" representation, every unit plays a role in the representation [78, 79, 27]. (The term "holographic" is in reference to holograms, which

encode light field information as interference patterns of reference waves and diffracted waves [80].)

If we have some understanding of the individual symbols such as pronunciation rules (e.g. vowel and consonant sounds), then we can even infer the written representations of words that we hear. This is impossible in a local representation such as the Chinese writing system—if we do not know a word, then there is little hope of inferring its symbol. This same idea also extends to inferring the meaning of new words based on understandings of root words and affixes. Thus, we can continue this iterative *coarse graining* to build distributed representations upon distributed representations.

Distributed representations are also robust to errors and exhibit *graceful degradation* in contrast to local representations. As a simple illustrative example, we can slhffue the leettrs in wdros, and we can still [hopefully] understand. Yes, we are technically cheating here since short operative words are preserved, and we can use context. However, this is precisely another example of a distributed representation, just at a higher level! In this case, the sentences are the "words", and the original words are now the "symbols" in the alphabet.

Perhaps a better example might be spelling errors or typos. Consider the word "representation". If we incorrectly spell it as "reprezentation", then most people will still be able to infer the intended word. We could even omit letters without making the word too unrecognizable: "reprezntation". Longer words are even more robust to errors: "antidisestablishmentarianism" versus "antidisestablshmentarianism". Can you find the mistake? In contrast, short words are less robust to errors even with the use of distributed representations. For example, the word "the" can become an entirely different word if we simply replace a single symbol: "tie". Thus, the dimensionality of a distributed representation affects robustness against errors.

3.3 High-Dimensional Distributed Representations

We now have the solution we were looking for at the beginning of this chapter—we can soften the SNR barrier on supply voltage scaling by using *high-dimensional distributed representations*. These representations provide us with *noise immunity* because they can continue to properly transmit information even in the presence of errors. From an information theory perspective, this should be of no surprise since this is simply a consequence of Shannon's theorem [81]. However, rather than using these high-dimensional representations solely as codes for error correction purposes (e.g. when communicating across a noisy channel), we are proposing to use them as the native representations themselves and directly compute with them. This would allow us to take advantage of the added computational efficiency provided by distributed representations as illustrated in the previous section.

Unfortunately, these representations also present a slew of new problems. High-dimensional representations require more bits per representation, resulting in larger word sizes and lower information rates. This will worsen the already problematic *von Neumann bottleneck* by increasing energy and bandwidth needed to transfer data between memory and the central

processing unit (CPU). Furthermore, if we use these representations to help reduce supply voltages, then circuits will run even slower due to smaller drive currents. Thus, to actually benefit from these representations, we will need in-memory logic—i.e. *processing-in-memory* (PIM)—to circumvent the von Neumann bottleneck. However, PIM has been proposed for decades [82] but has yet to meet widespread adoption due to various challenges [83]. Using different logic may alleviate some of these challenges but would change the kinds of computation that can be performed efficiently. Furthermore, it is still unclear at this point what kind of logic is needed to compute with these high-dimensional distributed representations. These issues will be addressed in the next chapter.

Chapter 4

Hyperdimensional Computing

4.1 Introduction

In the previous chapter, we discussed the benefits of using distributed representations over local representations. We found that distributed representations enable us to perform computations with more efficient space complexity and are robust against errors. Their robustness improves with increasing dimensionality, enabling them to exhibit graceful degradation. This makes the use of high-dimensional distributed representations a potential solution for reducing signal-to-noise ratio (SNR) requirements on supply voltages. However, these representations also have larger word sizes, which can severely increase energy consumption and limit performance due to data transfer between the central-processing unit (CPU) and memory (i.e. the *von Neumann bottleneck* [84]).

This problem can be solved by *processing-in-memory* (PIM) (i.e. using in-memory logic). However, this is a non-trivial solution—it has been proposed for decades [82] but has yet to meet widespread adoption due to various challenges [83]. For example, in-memory logic cannot efficiently access conventional CPU structures needed for efficient modern computing such as virtual memory and cache coherence mechanisms [85, 86, 83]. Thus, this would not necessarily work well for efficient general purpose computing. Furthermore, it is unclear what kind of logic is needed—how do we compute with high-dimensional distributed representations? What kinds of problems can we solve efficiently using these representations? Boolean algebra no longer applies, so we will need a new algebra.

Hyperdimensional computing is the answer. It is a cognitive computing model that relies on the systematic transformation of high-dimensional distributed representations [27]. Its operations are distributed and are consequently highly parallel, making it natural for performing energy efficient cognitive computing. It has already been successfully demonstrated to perform cognitive tasks such as language recognition and hand gesture recognition [28, 29, 87]. It performs one-shot learning [28, 29], eliminating the need for frequent weight updates commonly performed in conventional machine learning methods that rely on gradient descent. This alleviates the speed problem described in the previous chapter that arises from

reducing supply voltage.

With all the benefits that hyperdimensional computing purportedly brings, there is still one key link missing. While hyperdimensional computing provides well-defined operations for computing with high-dimensional distributed representations and also clearly defines how to construct these representations from random vectors, it does not define how to encode existing data into these representations. In practice, data is always measured in some sort of representation. A digital image, for example, is typically recorded as a raster image (i.e. a matrix of pixels). Thus, we need to find a way to map existing data representations to our desired high-dimensional distributed representations. This missing link is no fault of hyperdimensional computing, which is a complete framework for the scope it is defined for. We must simply bridge this gap before we can make use of this framework.

In this chapter, we propose to use higher-order spectra to bridge this gap. Higher-order spectra are natural candidates for encoding our desired representations due to their properties such as shift invariance, additivity, and homogeneity [88, 89, 90]. They capture nonlinear interactions [89, 91] and can be used to reconstruct the original data up to a single group operation [92]. Unfortunately, they are computationally expensive in both time and space. Even the third-order spectrum (i.e. the bispectrum)—which is the "lowest" higher-order spectrum—is too computationally expensive for practical widespread usage.

To address this problem, we discuss ways to approximate the bispectrum. This amounts to a low-rank approximation problem, which can be solved using singular value decomposition. However, this requires a full calculation of the bispectrum prior to approximation. Instead, we propose to approximate the bispectrum in one-shot without calculating it a priori. This is akin to a low-rank approximation with missing data (or a weighted low-rank approximation), which is an NP-hard problem [93]. We test our approach on a small dataset of music and successfully demonstrate one-shot unsupervised learning. Specifically, we demonstrate the following: (1) discovery of underlying structure; (2) song recognition; and (3) genre clustering.

4.2 Encoding with Random Vectors

Consider a p-dimensional vector of real-valued random variables

$$X^{\mathrm{T}} = (X_1, X_2, \ldots, X_p) \in \mathbb{R}^p \tag{4.1}$$

Note that the vector space can also be defined over a different field such as binary numbers $\{0, 1\}$ or complex numbers $\mathbb{C}$. If each random variable is independently and identically distributed (i.i.d.) with mean μ_1 and variance σ^2, then the expected squared norm of a random vector is given by

$$\mathrm{E}\,||X||^2 = \mathrm{E}\sum_{j=1}^{p} |X_j|^2 = p(\mu_1^2 + \sigma^2) \tag{4.2}$$

Thus, we expect random vectors to have length

$$||X|| \sim \sqrt{\mathrm{E}\,||X||^2} \sim \sqrt{p} \tag{4.3}$$

In other words, the tips of random vectors tend to reside on the surface of a sphere with radius $\sim \sqrt{p}$. The variance of the squared norm is

$$\begin{aligned}
\mathrm{Var}\,||X||^2 &= \mathrm{E}\,||X||^4 - (\mathrm{E}\,||X||^2)^2 \\
&= \mathrm{E}\left(\sum_{j=1}^{p} |X_j|^2\right)^2 - \left(\mathrm{E}\sum_{j=1}^{p} |X_j|^2\right)^2 \\
&= [p\mu_4 + (p^2-p)(\mu_1^2+\sigma^2)^2] - [p(\mu_1^2+\sigma^2)]^2 \\
&= p[\mu_4 - (\mu_1^2+\sigma^2)^2]
\end{aligned}$$

where $\mu_4 = \mathrm{E}\,X_j^4$ is the fourth moment of any of the i.i.d. random variables. Assuming the moments are finite, the fractional fluctuations of the squared norm are

$$\frac{\sqrt{\mathrm{Var}\,||X||^2}}{\mathrm{E}\,||X||^2} \sim \frac{1}{\sqrt{p}} \tag{4.4}$$

These fluctuations tend towards zero when the dimensionality p is high. Consequently, the random vectors are tightly concentrated near the surface of the sphere [94]:

$$||X||^2 \sim p \pm O(\sqrt{p}) \tag{4.5}$$

$$||X|| \sim \sqrt{p} \pm O(1) \tag{4.6}$$

In other words, if we draw random points, then they will lie near the surface of the sphere with high probability.

Let us now examine the angle between two random vectors $X, Y \in \mathbb{R}^p$. The expected squared inner product is

$$\begin{aligned}
\mathrm{E}(X^{\mathrm{T}}Y)^2 &= \mathrm{E}\left(\sum_{j=1}^{p} X_j Y_j\right)^2 \\
&= \mathrm{E}\sum_{j=1}^{p}\left(X_j^2 Y_j^2 + \sum_{k\neq j} X_j X_k Y_j Y_k\right) \\
&= \sum_{j=1}^{p}\left[\mathrm{E}(X_j^2)\,\mathrm{E}(Y_j^2) + \sum_{k\neq j} \mathrm{E}(X_j)\,\mathrm{E}(X_k)\,\mathrm{E}(Y_j)\,\mathrm{E}(Y_k)\right] \\
&= p(\mu_1^2+\sigma^2)^2 + (p^2-p)\mu_1^4
\end{aligned}$$

Combining this result with (4.2), we expect the cosine angle between two random vectors to tend towards

$$\cos\angle(X,Y) \sim \sqrt{\frac{\mathrm{E}(X^{\mathrm{T}}Y)^2}{\mathrm{E}\,||X||^2\,\mathrm{E}\,||Y||^2}} \sim \frac{1}{\sqrt{p}} \tag{4.7}$$

Consequently, as the dimensionality p increases, we expect random vectors to have $\cos \angle(X, Y) \sim 0$ or $\angle(X, Y) \sim \pi/2$. In contrast for two dimensions ($p = 2$), we would expect random vectors to have $\cos \angle(X, Y) \sim 1/\sqrt{2}$ or $\angle(X, Y) \sim \pi/4$. Thus, random vectors become approximately orthogonal in high dimensions.

In summary, (4.7) shows that random vectors in high dimensions are approximately orthogonal. This provides us with an exponential number of approximately orthogonal vectors that can be used to construct representations. These representations are fully distributed due to their i.i.d. elements. Furthermore, (4.3) and (4.4)-(4.6) show that these random vectors are concentrated near the surface of a sphere. This allows us to utilize various operations to systematically construct new distributed representations from existing representations. For example, computing the mean of a set of random vectors results in a mean-vector that is similar to its constituent vectors with statistical significance. Unitary transformations can also be performed to rotate vectors to other points on the sphere surface. These properties form the basis of hyperdimensional computing [27].

The above derivations were intentionally performed in a simple and intuitive way to illustrate the key concepts behind hyperdimensional computing, but there are more general concentration inequalities that can be proved rigorously [94]. The basic results above also apply to vector spaces defined over other fields such as the binary numbers and complex numbers. For binary vectors, the L_1 norm or *Hamming distance* is a more natural distance metric [27].

4.3 Higher-Order Statistics

Since hyperdimensional computing does not define a way to map existing data representations to high-dimensional distributed representations, we instead turn to higher-order statistics for their convenient properties. Consider a *random field* (a generalization of a random process), which is simply a collection of random variables

$$\{X_t : t \in T\} \tag{4.8}$$

where T is an arbitrary index set. If we are dealing with time-domain data, then the elements of the index set are time. If we are dealing with images, then the elements of the index set can be pixel coordinates.

The random field has a joint probability distribution

$$\Pr(\dots, X_t, \dots) \tag{4.9}$$

for all $t \in T$. The *characteristic function* is the Fourier transform of the probability distribution [95]:

$$\phi(t) = \mathrm{E}(e^{i \sum_{t \in T} t X_t}) \tag{4.10}$$

We can define n-th order *moments* in terms of the characteristic function [95]:

$$\mu_n(t_1, t_2, \dots, t_n) = (-i)^n \left. \frac{\partial^n \phi}{\partial t_1 \partial t_2 \cdots \partial t_n} \right|_0 \tag{4.11}$$

where $t_j \in T$ for $j = 1, 2, \ldots, n$. For order $n > 2$, these moments are considered higher-order moments or *higher-order statistics* [88, 89]. Note that there are multiple higher-order moments depending on the random variables selected by the indices.

Alternatively to moments, we can also define *cumulants* [95]:

$$\kappa_n(t_1, t_2, \ldots, t_n) = (-i)^n \left. \frac{\partial^n \ln(\phi)}{\partial t_1 \partial t_2 \cdots \partial t_n} \right|_0 \tag{4.12}$$

Notice that this equation is identical to the equation for moments except that it is defined in terms of the natural logarithm of the characteristic function. This provides cumulants with a few important properties that moments do not have [89, 90]:

- Shift Invariance: First-order cumulants are shift equivariant, but second-order and higher cumulants are shift *invariant*. Thus, a translation such as $X + c$ where c is a constant does not affect second-order and higher cumulants.
- Additivity: The cumulant of a sum of random variables is simply the sum of the individual cumulants. This can be useful when there are additive noise sources that have zero higher-order cumulants (e.g. Gaussian noise).
- Homogeneity: Scaling a random variable by a constant multiplies its cumulant by that same constant raised to the degree of homogeneity (i.e. the cumulant order). For example, if we scale a random variable X by a constant c, then its n-th order cumulant κ_n will become $c^n \kappa_n$. We can exploit this property later to attain some amount of scaling invariance.

We will be able to exploit some of these properties later. From (4.11) and (4.12), we can also see that moments are cumulants are related. In fact, moments can be expressed in terms of cumulants and vice versa using Bell polynomials, which have a profound combinatorial interpretation [96, 90, 97].

4.4 Higher-Order Spectra

Instead of attempting to directly estimate cumulants from data, it is often more computationally efficient to estimate *cumulant spectra* using the fast Fourier transform. Cumulant spectra are simply Fourier transforms of cumulant generating functions. Thus, the n-th order spectrum is given by the Fourier transform of the n-th order cumulant generating function:

$$K_n(f_1, f_2, \ldots, f_n) = \mathscr{F}\{\kappa_n(t_1, t_2, \ldots, t_n)\} \tag{4.13}$$

f_j are Fourier space indices conjugate to the original indices $t_j \in T$. Third-order or higher cumulant spectra are considered *higher-order spectra*. If the original index elements were points in time, then the conjugate indices are angular frequencies (note that we use f instead

of ω here because we later use ω for elements of an outcome set Ω). Similarly, if the original index elements were positions, then the conjugate indices are wavevectors.

As an example, let us consider the second-order spectrum. First, we compute the second-order cumulant generating function:

$$\begin{aligned}
\kappa_2(t_1, t_2) &= (-i)^2 \left.\frac{\partial^2 \ln(\phi)}{\partial t_1 \partial t_2}\right|_0 \\
&= (-i)^2 \left.\left(\frac{1}{\phi}\frac{\partial^2 \phi}{\partial t_1 \partial t_2} - \frac{1}{\phi^2}\frac{\partial \phi}{\partial t_1}\frac{\partial \phi}{\partial t_2}\right)\right|_0 \\
&= (-i)^2[\mathrm{E}(iX_{t_1} iX_{t_2}) - \mathrm{E}(iX_{t_1})\,\mathrm{E}(iX_{t_2})] \\
&= \mathrm{E}(X_{t_1} X_{t_2}) - \mathrm{E}(X_{t_1})\,\mathrm{E}(X_{t_2})
\end{aligned} \tag{4.14}$$

If we assume zero mean for simplicity, then this becomes

$$\kappa_2(t_1, t_2) = \mathrm{E}(X_{t_1} X_{t_2}) \tag{4.15}$$

Now we compute the Fourier transform. Recall that random variables are simply measurable functions between measurable spaces. For example, the random variable $X_t : \Omega \to S$ is a measurable function between a probability space $(\Omega, \mathcal{F}, P)$ and a measurable space $(S, \mathcal{S})$. The random variables are indexed by $t \in T$ but are only functions of outcomes $\omega \in \Omega$. More explicitly, the random variables can be written as functions $X(\omega; t)$ where ω is the variable and t is just an index parameter. Thus, we simply transform the indices:

$$\begin{aligned}
K_2(f_1, f_2) &= \mathscr{F}\{\kappa_2(t_1, t_2)\} \\
&= \mathscr{F} \int_\Omega X_{t_1}(\omega) X_{t_2}(\omega)\, dP \\
&= \int_\Omega X_{f_1}(\omega) X_{f_2}(\omega)\, dP \\
&= \mathrm{E}(X_{f_1} X_{f_2})
\end{aligned} \tag{4.16}$$

In other words, the second-order spectrum simply consists of second-order moments in Fourier space. If the random field is *homogeneous* (i.e. the joint distribution is *stationary*), then we lose a degree of freedom, and the *support* for the spectrum reduces. In this case, the spectrum is non-zero only when energy or momentum is conserved (this notion stems from physics where energy $\sim$ angular frequency and momentum $\sim$ wavevector):

$$\sum_j f_j = 0 \tag{4.17}$$

Thus, the spectrum reduces to

$$\begin{aligned}
K_2(f) &= \mathrm{E}(X_f X_{-f}) \\
&= \mathrm{E}(X_f \overline{X_f}) \\
&= \mathrm{E}(|X_f|^2)
\end{aligned} \tag{4.18}$$

This is the *power spectrum*.

Similarly for third-order, the cumulant generating function is

$$\begin{aligned}\kappa_3(t_1,t_2,t_3) = &\operatorname{E}(X_{t_1}X_{t_2}X_{t_3}) - \operatorname{E}(X_{t_1})\operatorname{E}(X_{t_2}X_{t_3}) - \operatorname{E}(X_{t_2})\operatorname{E}(X_{t_1}X_{t_3}) \\ &- \operatorname{E}(X_{t_3})\operatorname{E}(X_{t_1}X_{t_2}) + 2\operatorname{E}(X_{t_1})\operatorname{E}(X_{t_2})\operatorname{E}(X_{t_3})\end{aligned} \tag{4.19}$$

If we again assume zero mean, then this simplifies to

$$\kappa_3(t_1,t_2,t_3) = \operatorname{E}(X_{t_1}X_{t_2}X_{t_3}) \tag{4.20}$$

Taking the Fourier transform, we obtain the third-order spectrum

$$K_3(f_1,f_2,f_3) = \operatorname{E}(X_{f_1}X_{f_2}X_{f_3}) \tag{4.21}$$

This is the "lowest" higher-order spectrum. Confusingly, this particular spectrum is also known as the *bispectrum* because if we again assume a homogeneous field, then the support reduces, and we are only left with two degrees of freedom as per (4.17):

$$\begin{aligned}K_3(f_1,f_2) &= \operatorname{E}(X_{f_1}X_{f_2}X_{-(f_1+f_2)}) \\ &= \operatorname{E}(X_{f_1}X_{f_2}\overline{X_{f_1+f_2}})\end{aligned} \tag{4.22}$$

(Similarly, the fourth-order spectrum is referred to as the trispectrum.) Notice that unlike the power spectrum, the bispectrum retains phase information. This allows us to reconstruct original data from the bispectrum up to a single phase shift.

Let us now consider some of the issues with estimating higher-order spectra. In practice, we do not have perfect information about random variables and their distributions. Instead, we have *samples* of random variables. This means that taking the Fourier transform amounts to computing the discrete Fourier transform, and our frequency range is limited by sampling rate f_s. However, the usual Nyquist frequency $f_\mathrm{s}/2$ is effectively lower for higher-order spectra. For example, in (4.22) for the bispectrum, we see that the effective Nyquist frequency becomes $f_\mathrm{s}/4$ because we require sum frequencies $f_1 + f_2$. This reduction in Nyquist frequency worsens as the spectrum order increases, making it increasingly difficult to estimate higher order spectra.

To make matters worse, the sample size N of the transform limits the frequency resolution. We want to use large sample sizes for adequate frequency resolution, but this requires large memory space. For example, in (4.22) we see that the bispectrum of a homogeneous random field requires $O(N^2)$ elements. This may not seem like much from a space complexity perspective but can end up quite large in practice. As a concrete example, consider a music song sampled at 44.1 kHz (standard CD quality [98]). One second of this song contains 44100 samples. Thus, a Fourier transform of one second contains 22051 unique frequency bins due to the Nyquist frequency. However, as discussed above, the effective Nyquist frequency for the bispectrum reduces by another half. Consequently, the bispectrum of this one second requires the storage of 11026^2 elements or $\sim 10^8$ elements. If we assume 8 bytes per element

(4 bytes each for the real and imaginary parts), then this amounts to $\sim 10^9$ bytes or ~ 1 GB for just one second of music! The bispectrum of a full song (assuming a duration of four minutes or 240 seconds) would require 240^2 times more memory or ~ 60 TB. Furthermore, these calculations assumed a homogeneous random field, but a song is unlikely to have a stationary joint distribution. Thus, we would need to compute the full bispectrum in (4.21), which requires $O(N^3)$ elements.

This is clearly impractical and computationally infeasible—and this was only for computing a single bispectrum sample. Recall from (4.21) and (4.22) that the bispectrum is actually an expectation. Thus, even if we could calculate all of the elements described above, we would have to perform this calculation multiple times over different outcomes and then compute an ensemble average. It is likely for these computational challenges that the bispectrum has not experienced widespread usage. To make any practical usage of the bispectrum in hyperdimensional computing, we will need a way to efficiently approximate it.

4.5 Approximating the Bispectrum

Based on the Eckart-Young-Mirsky theorem, the best low-rank approximation of a matrix is determined from *singular value decomposition* (SVD) when "best" is measured by the Frobenius norm [99, 100]. The computational complexity of SVD is $O(N^3)$, which seems like a tractable problem from a computational complexity perspective but can still become practically difficult if n is large. Furthermore, computing the SVD requires the full matrix of elements. As discussed in the previous section, it is computationally infeasible to compute the bispectrum. Thus, we need a way to approximate the bispectrum without calculating it in full a priori. This amounts to a low-rank approximation problem with missing data (or, equivalently, a weighted low-rank approximation problem)—this is an NP-hard problem [93].

Fortunately, the bispectrum possesses special structure that we can exploit. Assuming a homogeneous random field, we can sketch out the structure of (4.22):

$$K_3 = \mathrm{E}\begin{pmatrix} X_{f_0}|X_{f_0}|^2 & & & & \\ X_{f_0}|X_{f_1}|^2 & X_{f_1}^2\overline{X_{f_2}} & & & \\ X_{f_0}|X_{f_2}|^2 & X_{f_1}X_{f_2}\overline{X_{f_3}} & X_{f_2}^2\overline{X_{f_4}} & & \\ X_{f_0}|X_{f_3}|^2 & X_{f_1}X_{f_3}\overline{X_{f_4}} & X_{f_2}X_{f_3}\overline{X_{f_5}} & X_{f_3}^2\overline{X_{f_6}} & \\ \vdots & & & & \ddots \end{pmatrix} \tag{4.23}$$

The bispectrum is symmetric, so we only show the elements in the lower triangle. Notice that different elements contain different amounts of information. For example, the first row/column contains the power spectrum. Therefore, each element of the first row/column effectively only contains amplitude information for a single frequency. In contrast, the diagonal elements each contain amplitude and phase information for two different frequencies.

Similarly, the subdiagonal elements each contain amplitude and phase information for three different frequencies. Thus, we can approximate the bispectrum by efficiently sampling the elements that are likely to contain more information (e.g. sampling the subdiagonal elements). Unfortunately, at the time of writing this thesis, this choice of elements is still mostly heuristic. We are currently working on a more formal and rigorous analysis for the future.

4.6 Example: One-Shot Unsupervised Learning on Music

In this section, we apply our bispectrum approximation approach together with hyperdimensional computing to demonstrate one-shot unsupervised learning on a small dataset of music. For simplicity, we convert songs from stereo to mono and downsample them to 22.05 kHz (half CD quality). Even with these parameters and our approximation method, it is still computationally difficult to estimate the bispectrum of a full song. Instead, we assume that a song can be decomposed into sections that are approximately locally stationary and ergodic. Thus, we compute the short-time Fourier transform (STFT) of a song $x[n]$ using a window function $w[n]$:

$$F_m[k] = \sum_{n=0}^{N-1} x[n+m]w[n]e^{-i2\pi kn/N} \tag{4.24}$$

$F_m[k]$ is the k-th Fourier coefficient of the song frame starting at sample m (note that we use F here instead of conventional notation X to prevent confusion with the random variables discussed in the previous sections); n is the sample number; and N is the window size (e.g. 22050 samples for one second). We assume the window function is positioned such that it is zero for $n < 0$ and $n > N - 1$. The remaining window behavior for $0 \leq n \leq N - 1$ is dependent on the specific choice of window function, and there are tradeoffs for choosing different window functions. The window size should be small enough to capture the smallest locally invariant features (e.g. if we want to capture features that have lengths on the order of seconds, then the window size should be on the order of seconds). Larger features are of less concern because they can be captured across multiple small frames. On the other hand, the window size should also be large enough to accurately estimate Fourier coefficients. (Alternatively, the estimated local bispectrum of a larger frame can be approximated by using a smaller window and mixing Fourier coefficients across multiple frames.) This conflict in window size and resolution suggests that a different basis such as wavelets might be more natural, but this is future work.

We store the collection of Fourier transforms $\{F_m[k]\}$ in a data matrix $\mathbf{X}$ where each row corresponds to the Fourier coefficients of a different frame. However, instead of computing the Fourier transform of every possible frame, we save on computation by only considering

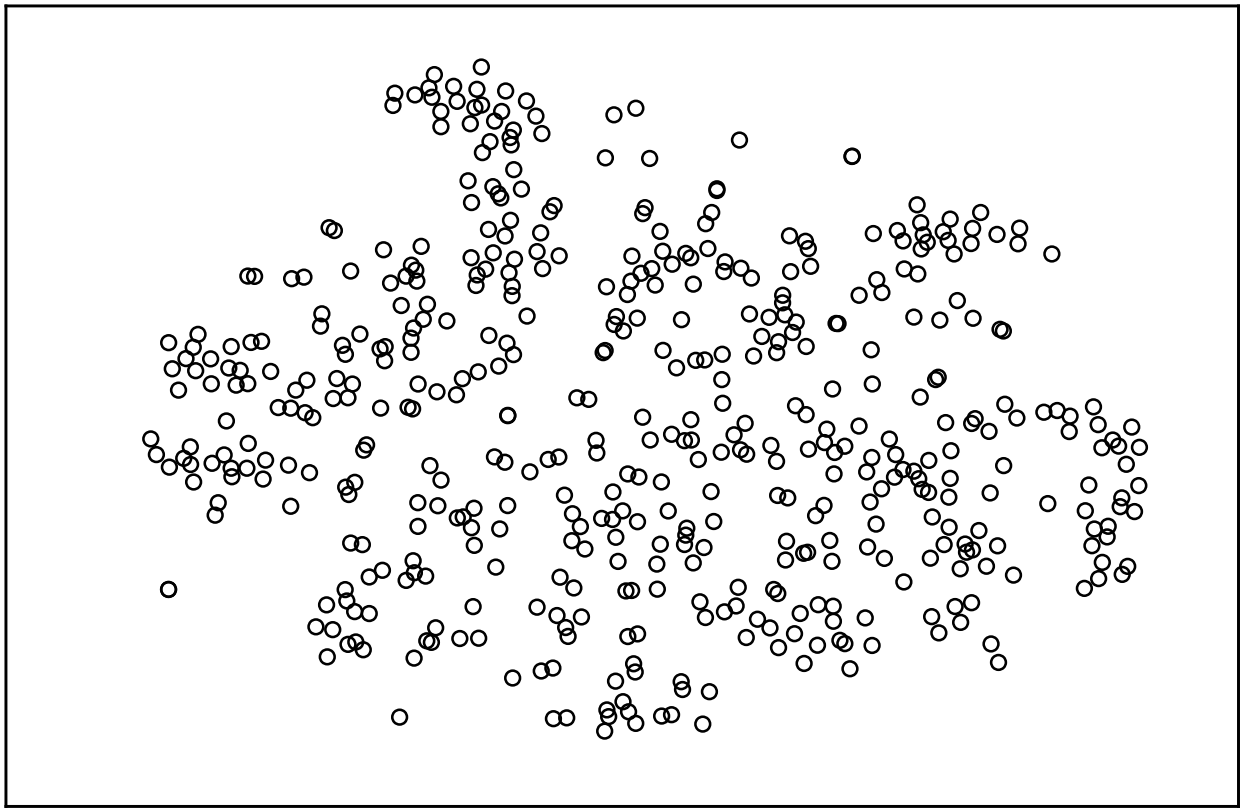

Figure 4.1: t-SNE visualization of the STFT of "Everything has Changed" by Taylor Swift featuring Ed Sheeran. There is some amount of clustering, but it is otherwise difficult to find any apparent intuitive structure here.

frames that are positioned at integer multiples of a constant shift Δm where $0 < \Delta m \leq N$:

$$\{F_m[k] : m \text{ is an integer multiple of } \Delta m\} \tag{4.25}$$

Thus, adjacent frames overlap by $N - \Delta m$ samples.

4.6.1 Discovering Underlying Structure

Let us now try to visualize any structure that might already be present in the STFT in our data matrix $\mathbf{X}$ by using a standard data visualization algorithm such as t-distributed stochastic neighbor embedding (t-SNE) [101]. An example visualization is shown in Fig. 4.1. Notice that there is no apparent structure other than some small amount of clustering. We will need to discover hidden structure by using unsupervised learning.

To accomplish this, we encode each feature vector in our data matrix into a high-dimensional distributed representation—a *hypervector*—using the subdiagonal of the bispectrum (4.22). There are likely better choices of elements to sample from the bispectrum,

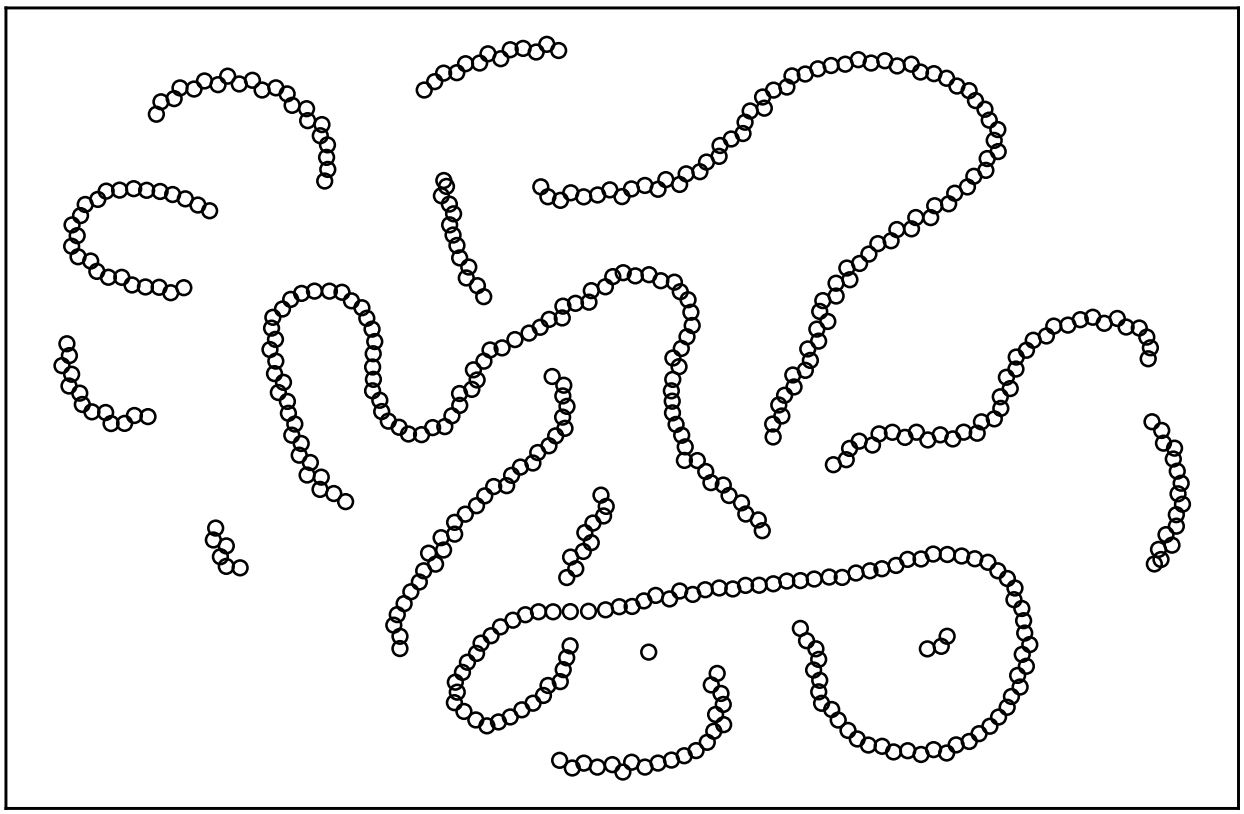

Figure 4.2: t-SNE visualization of the local bispectra estimates for "Everything has Changed" by Taylor Swift featuring Ed Sheeran. There is much more apparent structure here than in the original STFT shown in Fig. 4.1.

but we are still investigating this as discussed in Section 4.5. If we need amplitude scale invariance, then we can normalize the hypervectors at this point with respect to their norms to exploit cumulant homogeneity. Note that it can also help to follow up with a short moving average filter, e.g. five seconds, over the hypervectors to approximate ensemble averages assuming local ergodicity, but this essentially amounts to smoothing with a low-pass filter and is optional.

Thus, we obtain a new set of features that we can again visualize using t-SNE as shown in Fig. 4.6.1. Clearly, we have captured significantly more structure with our high-dimensional distributed representation than with the original STFT. Different points now cluster together in contiguous segments. It would be useful if we could understand where these features are coming from and why they might be clustering in this way. Fortunately, the operations we have performed so far are completely transparent, so we can determine the song frame that each point came from. Thus, we can label the points with approximate timestamps (remember that the points are not perfectly localized in time) as in Fig. 4.3. If we play

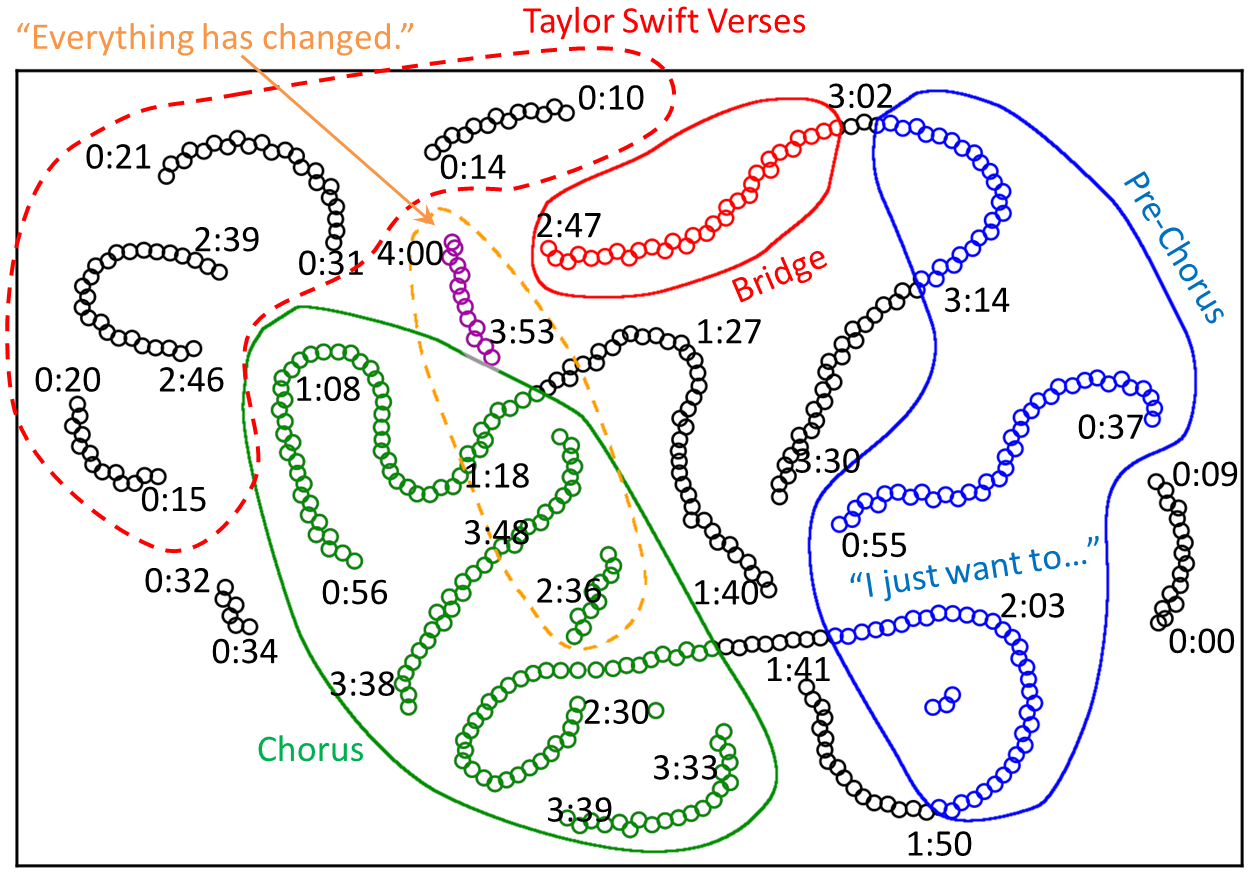

Figure 4.3: A "map" of the song "Everything has Changed" by Taylor Swift featuring Ed Sheeran. Data points locally cluster into contiguous segments that appear to intuitively capture different musical motifs. These segments also tend to cluster on a global scale to form regions that correspond to different parts of song structure (e.g. verses, the bridge, the chorus, etc.). The timestamps provide a sense of directionality due to the flow of time; different segments appear to flow correctly into different regions according to the structure of the song. However, not everything is correct (e.g. some points are oddly placed), but we are visualizing high-dimensional data in two dimensions—some structure will be lost. The timestamps are also approximate since the points are not perfectly localized in time.

through the song, then we can determine whether or not the structure we are looking at has any intuitive meaning. We find that contiguous segments appear to intuitively capture musical motifs. These segments also tend to cluster on a global scale to form different regions that correspond to different parts of a song (e.g. verses, the chorus, etc.). The segments have directionality due to the flow of time and appear to flow naturally from region to region with the flow of the song. Thus, in one shot—no gradient descent or other iterative optimization—we have uncovered a remarkable amount of structure in a completely unsupervised fashion.

4.6.2 Song Recognition

In addition to discovering underlying structure, we can also test our high-dimensional distributed representations in recognizing song slices that have never been seen before. Recall that our representations only encoded song frames of a fixed size with rigid alignment (e.g. every half second). In contrast, song slices in general can have arbitrary sizes and offsets. For example, at a sampling rate of 22.05 kHz, it is possible to have a frame shifted by 1 sample/22050 Hz $\approx 45.351\,\mu$s. Clearly, there are many more song slices of different sizes and offsets that we have not encoded. However, we should still be able to infer which song they come from.

To accomplish this, we first construct an overall representation of each song by composing together our existing representations. Assuming that our approximate bispectrum encoder provides us with sufficiently invariant and orthogonal high-dimensional distributed representations, then we can simply superpose these representations together (i.e. compute the mean-vector). This results in a single hypervector for each song. In other words, we represent each song as a collection of approximate locally invariant bispectra.

For testing, we draw random song slices of a given size from random start positions in a given song. We encode the slice using the same algorithm described above (i.e. estimate local bispectra from the STFT and then superpose the spectra together). Thus, each test slice is encoded as a single hypervector. We then measure the distance between this hypervector and all known song hypervectors. The song with the nearest hypervector is the predicted song.

We performed this test separately for ten classical songs, ten jazz songs, and a Taylor Swift album as shown in Fig. 4.4. Notice that recognition accuracy is lowest for short song slices, which simply do not have enough information to reliably distinguish between songs. Accuracy naturally improves for longer song slices. If we now test all of these songs together in a single experiment, then we might expect greater confusion among the songs. The result is shown in Fig. 4.5. Notice that there is minimal apparent degradation in accuracy, especially for longer song slices. The curves from Fig. 4.4 appear to simply overlay each other. However, a closer look reveals that there is a degradation for shorter song slices because of the greater possibility of confusion with more songs. This degradation is quite slight, demonstrating the quality of our approximate bispectra encoder.

To provide some perspective on these results, keep in mind that we are compressing songs by *orders of magnitude* in one shot. For reference, Table 4.6.2 compares the compression

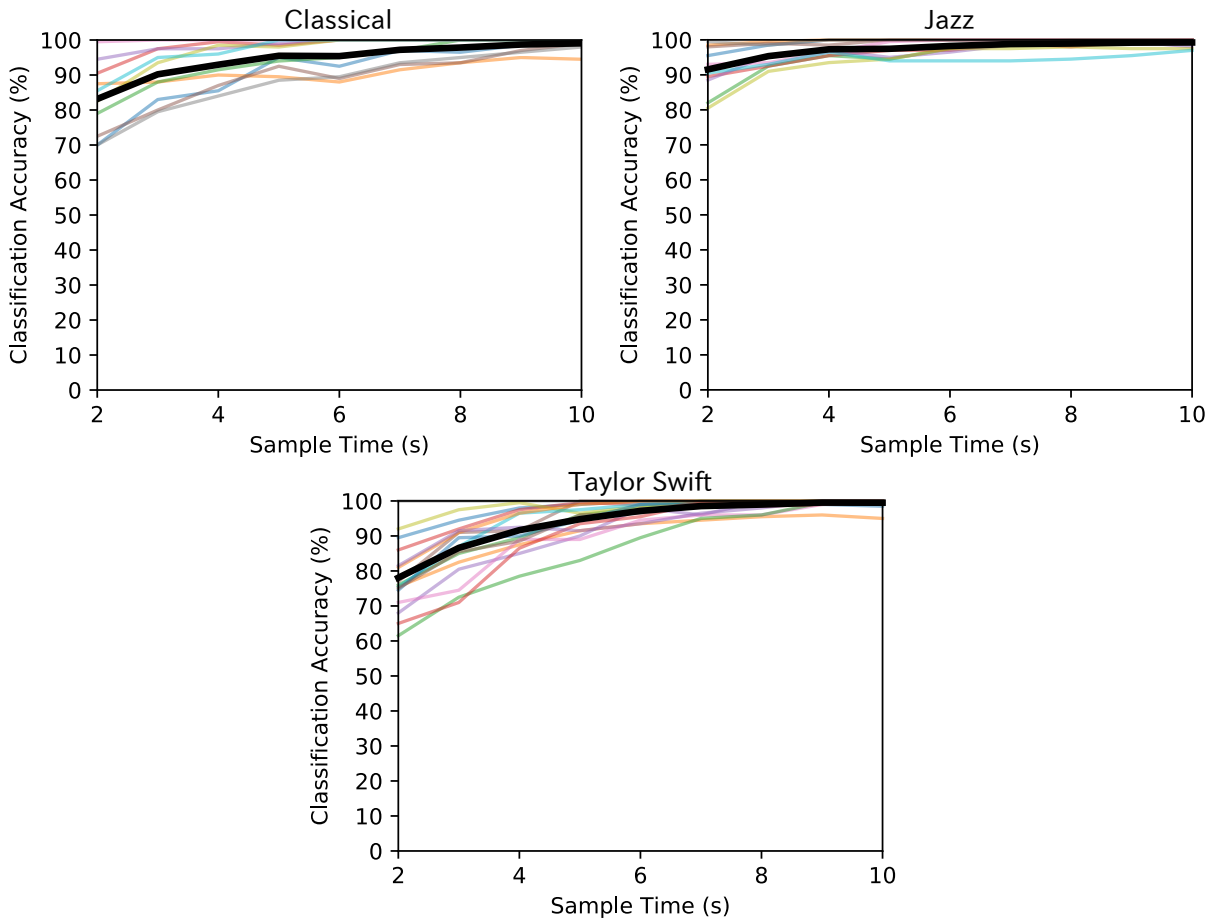

Figure 4.4: Song recognition accuracy for random slices of songs. Sample time refers to the duration of a song slice. Each colored curve corresponds to a particular song. The thick black curve is the mean performance averaged across all songs. Note that each plot is a separate experiment (e.g. we only test classical songs against classical songs in the "Classical" plot). Notice that accuracy improves with increasing sample time.

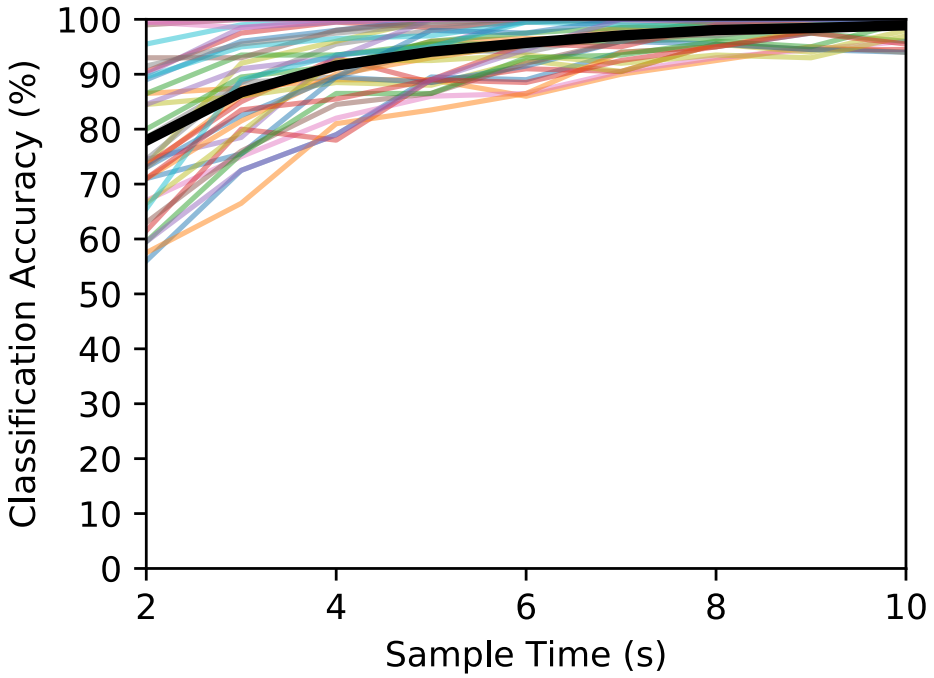

Figure 4.5: Song recognition accuracy for random slices of songs. All the songs from Fig. 4.4 are tested together here. Sample time refers to the duration of a song slice. Each colored curve corresponds to a particular song. The thick black curve is the mean performance averaged across all songs. Notice that accuracy improves with increasing sample time.

ratios (Original Memory Size / Compressed Memory Size) of MP3 against our hyperdimensional computing approach used here for the songs in the experiments above. MP3 was configured to use variable bitrate (with an average bitrate of 160 kbps) and could only achieve a compression ratio of ~ 7. In contrast, our high-dimensional distributed representations compress these songs by more than two orders of magnitude. Of course, our representations are highly distorted and cannot be used for music playback. However, we still retain enough information to recognize random song slices of different sizes.

Notice that longer songs are compressed by greater amounts because our hypervectors have a fixed size. Consequently, longer songs tend to perform worse because we are storing too much information into a single hypervector. This can be alleviated by simply using multiple hypervectors to represent longer songs. In fact, all of the curves in Fig. 4.5 increase in accuracy if we use more hypervectors per song. If we want to go beyond one shot learning, then we can even anneal these multiple hypervectors to better partition the local bispectra estimates to obtain a better representation.

Interestingly, if we use the diagonal of the bispectrum instead of the subdiagonal, then all of the curves in Fig. 4.5 lower in accuracy. Similarly, if we use the power spectrum

Table 4.1: Compression ratios for MP3 versus hyperdimensional (HD) computing.

		Compression Ratio	
Song	Size (kB)	MP3	HD
Beethoven - 5th Symphony	21537.0	6.799	250.044
Beethoven - Fur Elise	7544.3	8.424	87.589
Chopin - Ballade No. 1 in G minor	23600.4	7.008	274.000
Chopin - Nocturne No. 2 in E-flat Major, Op. 9, No. 2	10381.5	7.503	120.529
Hall of the Mountain King	6588.1	7.292	76.488
Liszt - Hungarian Rhapsody No.2	20299.5	6.344	235.677
Mozart - Turkish March	9129.4	6.735	105.992
Rachmaninov - Piano Concerto in C Minor	28986.8	6.779	336.536
Richard Wagner - Ride Of The Valkyries	12926.3	9.104	150.074
The Nutcracker - March	5630.3	7.111	65.368
Billy Strayhorn - Chelsea Bridge	8530.9	6.045	99.044
Bob Haggart (John Coltrane) - What's New?	9744.8	7.150	113.137
Bob Haggart - What's New?	12382.9	7.374	143.765
Chick Corea (Steve Gadd) - Windows	27873.0	6.728	323.605
Chick Corea - Mirror, Mirror	16176.4	7.343	187.808
Chick Corea - Windows	8300.3	7.237	96.366
Duke Ellington - Sophisticated Lady	7962.8	8.131	92.448
John Coltrane - Giant Steps	12352.5	7.254	143.413
Miles Davis - Blue in Green	14540.7	7.577	168.817
Thelonious Monk - Evidence	22632.8	7.370	262.766
Taylor Swift - State of Grace	12735.6	6.967	147.860
Taylor Swift - Red	9607.8	7.192	111.547
Taylor Swift - Treacherous	10454.8	7.084	121.380
Taylor Swift - I Knew You Were Trouble	9462.6	7.021	109.860
Taylor Swift - All Too Well	14175.7	7.160	164.580
Taylor Swift - 22	9996.6	7.267	116.060
Taylor Swift - I Almost Do	10532.3	7.259	122.280
Taylor Swift - We Are Never Ever Getting Back Together	8318.1	7.247	96.573
Taylor Swift - Stay Stay Stay	8865.4	6.759	102.927
Taylor Swift - The Last Time (feat. Gary Lightbody)	12877.4	7.335	149.507
Taylor Swift - Holy Ground	8740.8	6.962	101.480
Taylor Swift - Sad Beautiful Tragic	12268.2	7.557	142.433
Taylor Swift - The Lucky One	10347.4	7.206	120.133
Taylor Swift - Everything has Changed (feat. Ed Sheeran)	10569.1	7.105	122.707
Taylor Swift - Starlight	9500.4	7.222	110.300
Taylor Swift - Begin Again	10233.2	7.128	118.807

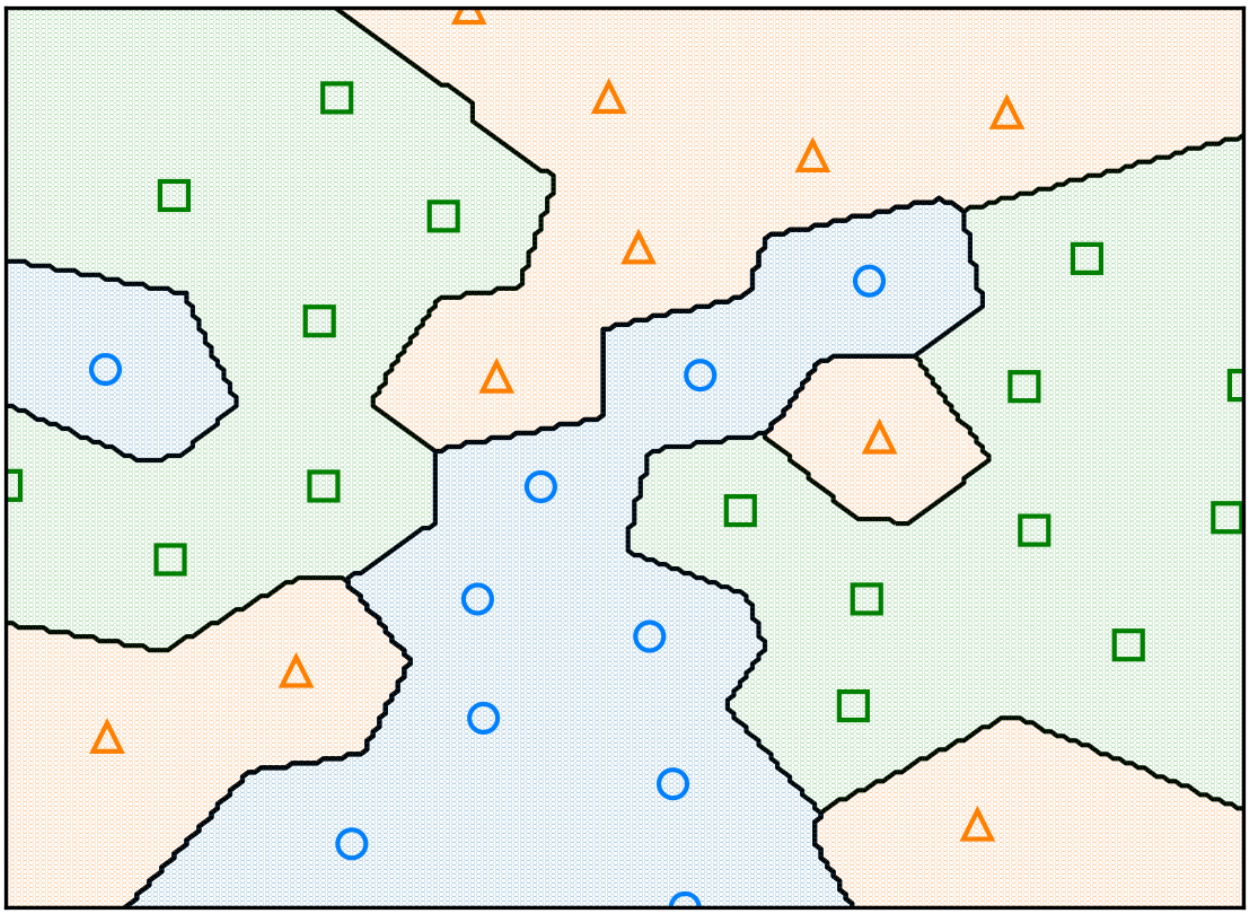

Figure 4.6: t-SNE visualization of song hypervectors. Songs from the same genre/artist tend to cluster together.

instead of the bispectrum, then we observe an even further degradation in accuracy. In contrast, sampling more elements from the bispectrum will improve performance since our approximation of the bispectrum will improve. This heuristically confirms our discussion in Section 4.5 regarding choosing elements with more information.

4.6.3 Genre Clustering

Now that we have composed representations of overall songs, we can see if our representations have uncovered any apparent structure at the song level. If we treat the song hypervectors as feature vectors, then we can use t-SNE to visualize their structure as shown in Fig. 4.6. Notice that songs from the same genre (or Taylor Swift) tend to cluster together. Unfortunately, we do not have many songs here (it turns out that music is a poor choice for data due to copyright issues). Instead, we can use 30 second song clips from analogous genres (replace Taylor Swift with pop) in the George Tzanetakis (GTZAN) dataset. The

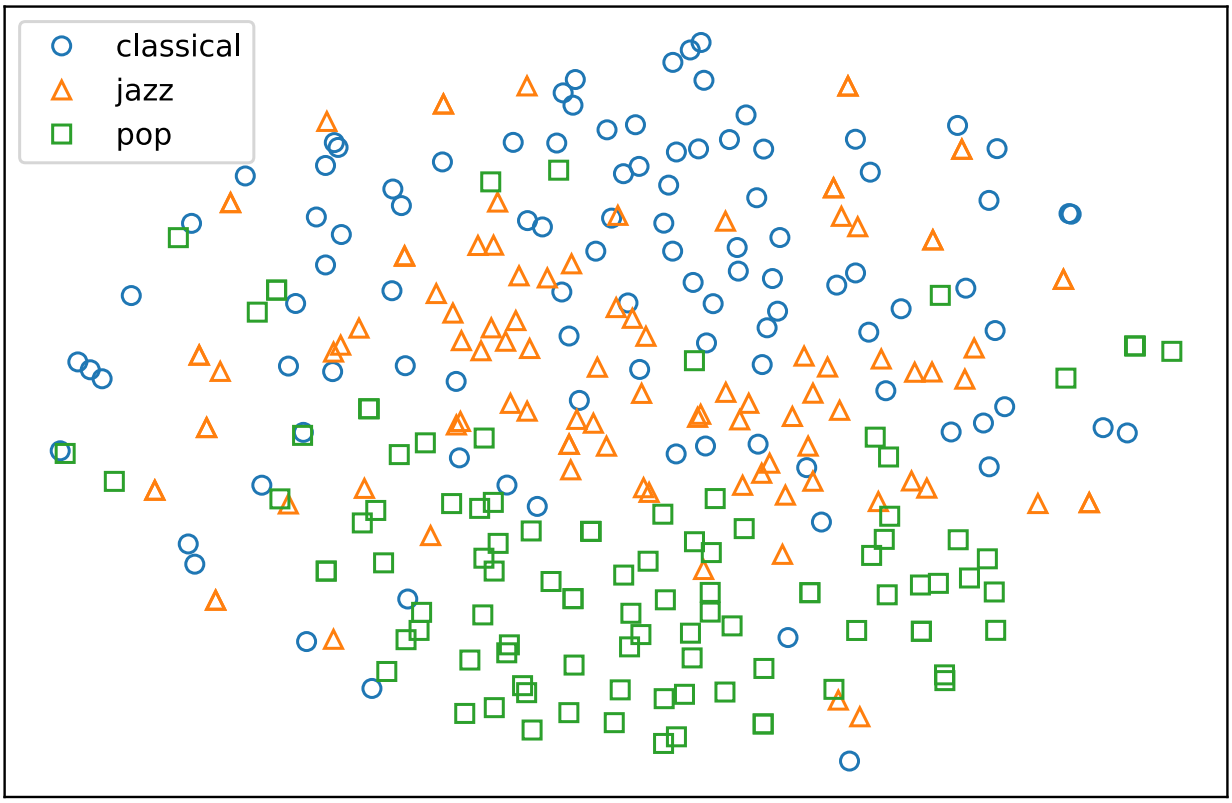

Figure 4.7: t-SNE visualization of song hypervectors encoded from the George Tzanetakis (GTZAN) dataset. Our high-dimensional distributed representations result in some amount of natural genre clustering.

visualization of the corresponding song hypervectors is shown in Fig. 4.7. Notice that the genres tend to cluster together, and our representations uncovered this structure through one-shot unsupervised learning. At this point, we can easily add a multilayer perceptron to perform genre classification using our song representations if we so desired. However, genre classification is actually an ill-posed problem because music genres are not well-defined [102, 103, 104, 105]. Nonetheless, the use of music in this entire section was simply meant to serve as an illustrative example.

Chapter 5

Conclusion

In conclusion, we were able to fill in some of the theoretical gaps described in Chapter 1, but there is still much work left to do:

- In Chapter 2, we demystified ferroelectric negative capacitance by constructing a microscopic model to illustrate how negative capacitance arises and can be stabilized in a prototypical displacive ferroelectric. We found that the phenomenological model used to originally derive ferroelectric negative capacitance can have correspondences with physical crystal configurations. Thus, the S curve that some believe is a theoretical artifact is in fact quite real.

  However, in our model, we used a mean-field approximation, which essentially results in a single domain ferroelectric. In practice, ferroelectrics often break into multiple domains, which can significantly complicate the dipolar fields. This many-body problem can be solved using simulation, but then we lose physical intuition. Alternatively, it would be more illuminating to model ferroelectric negative capacitance using renormalization group theory and examine order parameter fluctuations when the negative capacitance is stabilized by depolarization fields.

- Additionally in Chapter 2, we analyzed energy flow in a ferroelectric-dielectric system from the perspective of Poynting's theorem. We showed that the ferroelectric supplies extra energy to the dielectric during charging by transferring the energy stored in its phase transition. This energy is transferred back to the ferroelectric from the dielectric during discharging. Thus, there is no extra energy needed from an external source because the ferroelectric's phase transition energy is recycled, and the overall energy dissipation is reduced below the $^1/_2\,CV^2$ limit during charging and discharging.

  However, similarly to the microscopic model, we assumed a single domain ferroelectric. A multidomain ferroelectric will have a more complex pattern of energy flow due to more complicated dipolar fields. We do not believe this will affect the overall energy dissipation, but it still might be worth investigating.

- In Chapters 3 and 4, we discussed how computing with high-dimensional distributed representations is natural for energy efficient cognitive computing. However, the actual hardware implementation details play a key role in determining whether or not we actually save any energy. The high-dimensional representations will require in-memory logic to avoid the von Neumann bottleneck, but this logic can be difficult to design depending on the logic needed. At this point in time, it is still not entirely clear what exact operations are needed to perform non-trivial cognitive tasks using hyperdimensional computing. We need to continue studying the mathematics of hyperdimensional computing to better understand the hardware requirements.

- In Chapter 4, we proposed to use an approximation of the bispectrum to encode data into high-dimensional distributed representations. Using this approach together with hyperdimensional computing, we were able to successfully perform one-shot unsupervised learning on a small dataset of music for demonstration purposes.

  However, finding a low-rank approximation of the bispectrum without calculating it in full a priori is an NP-hard problem. Instead, we discussed ways to approximate the bispectrum in one-shot by sampling specific elements based on the structure of the bispectrum. As an example, we suggested to sample the subdiagonal elements. However, this choice is still largely heuristic, and there are likely better choices. This problem must be formalized and analyzed more rigorously to provide us with a better understanding of which elements of the bispectrum to sample.

Bibliography

[1] Inc. Gartner. *Gartner Says Worldwide Sales of Smartphones Recorded First Ever Decline During the Fourth Quarter of 2017*. Feb. 2018. URL: https://www.gartner.com/en/newsroom/press-releases/2018-02-22-gartner-says-worldwide-sales-of-smartphones-recorded-first-ever-decline-during-the-fourth-quarter-of-2017.

[2] Statista. *Smartphone penetration worldwide 2014-2021*. Dec. 2017. URL: https://www.statista.com/statistics/203734/global-smartphone-penetration-per-capita-since-2005/.

[3] Jacob Poushter. *Smartphone Ownership and Internet Usage Continues to Climb in Emerging Economies*. Feb. 2016. URL: http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/.

[4] Statista. *IoT: number of connected devices worldwide 2012-2025*. Nov. 2016. URL: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/.

[5] Roger Parloff. *The AI Revolution: Why Deep Learning is Suddenly Changing Your Life*. Sept. 2016. URL: http://fortune.com/ai-artificial-intelligence-deep-machine-learning/.

[6] Inc. CB Information Services. *Get Smart: 13 Industries Using Deep Learning To Make Huge Leaps Forward*. Aug. 2016. URL: https://www.cbinsights.com/research/industries-disrupted-deep-learning/.

[7] Michael Chui et al. *Notes from the AI Frontier: Insights from Hundreds of Use Cases*. Tech. rep. 2018.

[8] Bernard Marr. *The 4th Industrial Revolution: How Mining Companies Are Using AI, Machine Learning And Robots*. Sept. 2018. URL: https://www.forbes.com/sites/bernardmarr/2018/09/07/the-4th-industrial-revolution-how-mining-companies-are-using-ai-machine-learning-and-robots/#4268b4f7497e.

[9] J. Tang et al. "Enabling Deep Learning on IoT Devices". In: *Computer* 50.10 (2017), pp. 92–96. ISSN: 0018-9162. DOI: 10.1109/MC.2017.3641648.

[10] H. Li, K. Ota, and M. Dong. "Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing". In: *IEEE Network* 32.1 (Jan. 2018), pp. 96–101. ISSN: 0890-8044. DOI: 10.1109/MNET.2018.1700202.

[11] M. Mohammadi et al. "Deep Learning for IoT Big Data and Streaming Analytics: A Survey". In: *IEEE Communications Surveys Tutorials* 20.4 (Fourthquarter 2018), pp. 2923–2960. ISSN: 1553-877X. DOI: 10.1109/COMST.2018.2844341.

[12] Jesus Rodriguez. *What's New in Deep Learning Research: Mobile Deep Learning with Google MnasNet*. Aug. 2018. URL: https://towardsdatascience.com/whats-new-in-deep-learning-research-mobile-deep-learning-with-google-mnasnet-cf9844d30ae8.

[13] Adam Coates et al. "Deep Learning with COTS HPC Systems". In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML'13. Atlanta, GA, USA: JMLR.org, 2013, pp. III-1337–III-1345. URL: http://dl.acm.org/citation.cfm?id=3042817.3043086.

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.

[15] Mingsheng Long et al. "Deep Transfer Learning with Joint Adaptation Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, June 2017, pp. 2208–2217. URL: http://proceedings.mlr.press/v70/long17a.html.

[16] Simon M. Sze and Kwok K. Ng. *Physics of Semiconductor Devices*. 3rd ed. Wiley-Interscience, Oct. 2006.

[17] Yuan Taur and Tak H. Ning. *Fundamentals of Modern VLSI Devices*. 2nd ed. Cambridge University Press, June 2013.

[18] Sayeef Salahuddin and Supriyo Datta. "Use of Negative Capacitance to Provide Voltage Amplification for Low Power Nanoscale Devices". In: *Nano Letters*. Nano Lett. (USA) 8.2 (Feb. 2008), pp. 405–410. ISSN: 1530-6984. DOI: 10.1021/nl071804g. URL: http://pubs.acs.org/doi/abs/10.1021/nl071804g.

[19] Daniel J R Appleby et al. "Experimental observation of negative capacitance in ferroelectrics at room temperature". In: *Nano Letters* 14.7 (July 2014), pp. 3864–3868. ISSN: 15306992. DOI: 10.1021/nl5017255. URL: http://www.ncbi.nlm.nih.gov/pubmed/24915057%20http://pubs.acs.org/doi/abs/10.1021/nl5017255.

[20] Asif Islam Khan et al. "Experimental evidence of ferroelectric negative capacitance in nanoscale heterostructures". In: *Applied Physics Letters* 99.11 (Sept. 2011), p. 113501. ISSN: 0003-6951. DOI: 10.1063/1.3634072. arXiv: 1103.4419. URL: http://arxiv.org/abs/1103.4419%20http://link.aip.org/link/APPLAB/v99/i11/p113501/s1%7B%5C&%7DAgg=doi%20http://aip.scitation.org/doi/10.1063/1.3634072.

[21] Asif Islam Khan et al. "Negative capacitance in a ferroelectric capacitor". In: *Nature Materials* 14.2 (Dec. 2014), pp. 182–186. ISSN: 1476-1122. DOI: 10.1038/nmat4148. URL: http://www.nature.com/doifinder/10.1038/nmat4148.

[22] Asif Islam Khan et al. "Negative Capacitance Behavior in a Leaky Ferroelectric". In: *IEEE Transactions on Electron Devices* 63.11 (Nov. 2016), pp. 4416–4422. ISSN: 0018-9383. DOI: 10.1109/TED.2016.2612656. URL: http://ieeexplore.ieee.org/document/7582465/.

[23] Pavlo Zubko et al. "Negative capacitance in multidomain ferroelectric superlattices". In: *Nature* (June 2016), pp. 1–15. ISSN: 0028-0836. DOI: 10.1038/nature17659. URL: http://www.nature.com/doifinder/10.1038/nature17659.

[24] D. J. Frank et al. "Device scaling limits of Si MOSFETs and their application dependencies". In: *Proceedings of the IEEE* 89.3 (Mar. 2001), pp. 259–288. ISSN: 0018-9219. DOI: 10.1109/5.915374.

[25] Jan M. Rabaey, Anantha Chandrakasan, and Borivoje Nikolic. *Digital Integrated Circuits: A Design Perspective*. 2nd ed. Pearson, Jan. 2003.

[26] Seyed Mohammad Seyedzadeh et al. "Mitigating Bitline Crosstalk Noise in DRAM Memories". In: *Proceedings of the International Symposium on Memory Systems*. MEMSYS '17. Alexandria, Virginia: ACM, 2017, pp. 205–216. ISBN: 978-1-4503-5335-9. DOI: 10.1145/3132402.3132410. URL: http://doi.acm.org/10.1145/3132402.3132410.

[27] Pentti Kanerva. "Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors". In: *Cognitive Computation* 1.2 (June 2009), pp. 139–159. ISSN: 1866-9956. DOI: 10.1007/s12559-009-9009-8. URL: http://link.springer.com/10.1007/s12559-009-9009-8.

[28] Abbas Rahimi et al. "Hyperdimensional biosignal processing: A case study for EMG-based hand gesture recognition". In: *2016 IEEE International Conference on Rebooting Computing (ICRC)*. IEEE, Oct. 2016, pp. 1–8. ISBN: 978-1-5090-1370-8. DOI: 10.1109/ICRC.2016.7738683. URL: http://ieeexplore.ieee.org/document/7738683/.

[29] Abbas Rahimi, Pentti Kanerva, and Jan M. Rabaey. "A Robust and Energy-Efficient Classifier Using Brain-Inspired Hyperdimensional Computing". In: *Proceedings of the 2016 International Symposium on Low Power Electronics and Design - ISLPED '16* (2016), pp. 64–69. DOI: 10.1145/2934583.2934624. URL: http://dl.acm.org/citation.cfm?doid=2934583.2934624.

[30] R. H. Dennard et al. "Design of ion-implanted MOSFET's with very small physical dimensions". In: *IEEE Journal of Solid-State Circuits* 9.5 (Oct. 1974), pp. 256–268. ISSN: 0018-9200. DOI: 10.1109/JSSC.1974.1050511.

[31] Robert F. Pierret. *Semiconductor Device Fundamentals*. 2nd ed. Addison Wesley, Apr. 1996.

[32] G. Baccarani, M. R. Wordeman, and R. H. Dennard. "Generalized scaling theory and its application to a $^1/_4$ micrometer MOSFET design". In: *IEEE Transactions on Electron Devices* 31.4 (Apr. 1984), pp. 452–462. ISSN: 0018-9383. DOI: 10.1109/T-ED.1984.21550.

[33] *International Roadmap for Devices and Systems 2017 Edition*. 2017.

[34] *International Technology Roadmap for Semiconductors 2001 Edition*. 2001.

[35] K. Mistry et al. "A 45nm Logic Technology with High-k+Metal Gate Transistors, Strained Silicon, 9 Cu Interconnect Layers, 193nm Dry Patterning, and 100% Pb-free Packaging". In: *2007 IEEE International Electron Devices Meeting*. Dec. 2007, pp. 247–250. DOI: 10.1109/IEDM.2007.4418914.

[36] J. L. Hoyt et al. "Strained silicon MOSFET technology". In: *Digest. International Electron Devices Meeting,* Dec. 2002, pp. 23–26. DOI: 10.1109/IEDM.2002.1175770.

[37] S. Thompson et al. "A 90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 /spl mu/m/sup 2/ SRAM cell". In: *Digest. International Electron Devices Meeting,* Dec. 2002, pp. 61–64. DOI: 10.1109/IEDM.2002.1175779.

[38] T. Ghani et al. "A 90nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors". In: *IEEE International Electron Devices Meeting 2003*. Dec. 2003, pp. 11.6.1–11.6.3. DOI: 10.1109/IEDM.2003.1269442.

[39] S. E. Thompson et al. "A logic nanotechnology featuring strained-silicon". In: *IEEE Electron Device Letters* 25.4 (Apr. 2004), pp. 191–193. ISSN: 0741-3106. DOI: 10.1109/LED.2004.825195.

[40] S. E. Thompson et al. "A 90-nm logic technology featuring strained-silicon". In: *IEEE Transactions on Electron Devices* 51.11 (Nov. 2004), pp. 1790–1797. ISSN: 0018-9383. DOI: 10.1109/TED.2004.836648.

[41] D. Hisamoto, T. Kaga, and E. Takeda. "Impact of the vertical SOI 'DELTA' structure on planar device technology". In: *IEEE Transactions on Electron Devices* 38.6 (June 1991), pp. 1419–1424. ISSN: 0018-9383. DOI: 10.1109/16.81634.

[42] D. Hisamoto et al. "FinFET-a self-aligned double-gate MOSFET scalable to 20 nm". In: *IEEE Transactions on Electron Devices* 47.12 (Dec. 2000), pp. 2320–2325. ISSN: 0018-9383. DOI: 10.1109/16.887014.

[43] Victor V. Zhirnov and Ralph K. Cavin. "Negative capacitance to the rescue?" In: *Nature Nanotechnology* 3.2 (Feb. 2008), pp. 77–78. ISSN: 1748-3387. DOI: 10.1038/nnano.2008.18. arXiv: 0707.2073. URL: http://www.nature.com/articles/nnano.2008.18.

[44] Thomas N Theis and Paul M Solomon. "It's Time to Reinvent the Transistor!" In: *Science* 327.5973 (Mar. 2010), pp. 1600–1601. ISSN: 0036-8075. DOI: 10.1126/science.1187597. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1187597.

[45] Neil W. Ashcroft and N. David Mermin. *Solid State Physics*. 1st ed. Cengage Learning, Jan. 1976.

[46] John Michael Ziman. *Principles of the Theory of Solids*. 2nd ed. Cambridge University Press, Nov. 1979.

[47] Charles Kittel. *Introduction to Solid State Physics*. 8th ed. Wiley, Nov. 2004.

[48] Charles Kittel and Herbert Kroemer. *Thermal Physics*. 2nd ed. W. H. Freeman, Jan. 1980.

[49] Stanley Wolf. *The Submicron MOSFET*. Vol. 3. Silicon Processing for the VLSI Era. Lattice Press, Oct. 1994.

[50] A. C. Seabaugh and Q. Zhang. "Low-Voltage Tunnel Transistors for Beyond CMOS Logic". In: *Proceedings of the IEEE* 98.12 (Dec. 2010), pp. 2095–2110. ISSN: 0018-9219. DOI: 10.1109/JPROC.2010.2070470.

[51] S. Agarwal and E. Yablonovitch. "Why tunneling FETs don't work, and how to fix it". In: *2013 Third Berkeley Symposium on Energy Efficient Electronic Systems (E3S)*. Oct. 2013, pp. 1–2. DOI: 10.1109/E3S.2013.6705868.

[52] K. K. Young. "Short-channel effect in fully depleted SOI MOSFETs". In: *IEEE Transactions on Electron Devices* 36.2 (Feb. 1989), pp. 399–402. ISSN: 0018-9383. DOI: 10.1109/16.19942.

[53] Richard P. Feynman, Robert B. Leighton, and Matthew Sands. *Mainly Electromagnetism and Matter*. New Millennium. Vol. 2. The Feynman Lectures on Physics. Basic Books, Oct. 2011.

[54] Justin C. Wong and Sayeef Salahuddin. "Can piezoelectricity lead to negative capacitance?" In: *2014 IEEE International Electron Devices Meeting*. 2014, pp. 13.5.1–13.5.4. ISBN: 978-1-4799-8001-7. DOI: 10.1109/IEDM.2014.7047046. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7047046.

[55] Justin Wong and Sayeef Salahuddin. "Piezoelectric Negative Capacitance". MA thesis. EECS Department, University of California, Berkeley, May 2015. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2015/EECS-2015-57.html.

[56] L. D. Landau. "On the theory of phase transitions". In: *Zh. Eksp. Teor. Fiz.* 7 (1937). [Ukr. J. Phys.53,25(2008)], pp. 19–32.

[57] Karin M. Rabe, Charles H. Ahn, and Jean-Marc Triscone. *Physics of Ferroelectrics*. Springer-Verlag Berlin Heidelberg, 2007.

[58] B.R.A. Nijboer and F.W. De Wette. "The internal field in dipole lattices". In: *Physica* 24.1-5 (Jan. 1958), pp. 422–431. ISSN: 00318914. DOI: 10.1016/S0031-8914(58)95803-8. URL: http://linkinghub.elsevier.com/retrieve/pii/S0031891458958038.

[59] J. M. Luttinger and L. Tisza. "Theory of Dipole Interaction in Crystals". In: *Physical Review* 70.11-12 (Dec. 1946), pp. 954–964. ISSN: 0031-899X. DOI: 10.1103/PhysRev.70.954. URL: https://link.aps.org/doi/10.1103/PhysRev.70.954.

[60] J. C. Slater. "The lorentz correction in barium titanate". In: *Physical Review* 78.6 (June 1950), pp. 748–761. ISSN: 0031899X. DOI: 10.1103/PhysRev.78.748. URL: https://link.aps.org/doi/10.1103/PhysRev.78.748.

[61] R. R. Mehta, B. D. Silverman, and J. T. Jacobs. "Depolarization fields in thin ferroelectric films". In: *Journal of Applied Physics* 44.8 (Aug. 1973), pp. 3379–3385. ISSN: 0021-8979. DOI: 10.1063/1.1662770. URL: http://aip.scitation.org/doi/10.1063/1.1662770.

[62] A. M. Bratkovsky and A. P. Levanyuk. "Very large dielectric response of thin ferroelectric films with the dead layers". In: *Physical Review B* 63.13 (Mar. 2001), p. 132103. ISSN: 0163-1829. DOI: 10.1103/PhysRevB.63.132103. arXiv: 0011005 [cond-mat]. URL: http://arxiv.org/abs/cond-mat/0011005%20http://link.aps.org/doi/10.1103/PhysRevB.63.132103%20https://link.aps.org/doi/10.1103/PhysRevB.63.132103.

[63] A. M. Bratkovsky and A. P. Levanyuk. "Depolarizing field and "real" hysteresis loops in nanometer-scale ferroelectric films". In: *Applied Physics Letters* 89.25 (2006), p. 253108. ISSN: 00036951. DOI: 10.1063/1.2408650. URL: http://scitation.aip.org/content/aip/journal/apl/89/25/10.1063/1.2408650.

[64] C. B. Eom et al. "Fabrication and properties of epitaxial ferroelectric heterostructures with (SrRuO 3) isotropic metallic oxide electrodes". In: *Applied Physics Letters* 63.18 (Nov. 1993), pp. 2570–2572. ISSN: 0003-6951. DOI: 10.1063/1.110436. URL: http://aip.scitation.org/doi/10.1063/1.110436.

[65] Sayeef Salahuddin and Supriyo Datta. "Interacting systems for self-correcting low power switching". In: *Applied Physics Letters* 90.9 (Feb. 2007), p. 093503. ISSN: 0003-6951. DOI: 10.1063/1.2709640. arXiv: 0611569v1 [arXiv:cond-mat]. URL: http://aip.scitation.org/doi/10.1063/1.2709640.

[66] Alexandru Rusu et al. "Metal-Ferroelectric-Meta-Oxide-semiconductor field effect transistor with sub-60mV/decade subthreshold swing and internal voltage amplification". In: *2010 International Electron Devices Meeting*. IEEE, Dec. 2010, pp. 16.3.1–16.3.4. ISBN: 978-1-4424-7418-5. DOI: 10.1109/IEDM.2010.5703374. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5703374%20http://ieeexplore.ieee.org/xpls/abs%7B%5C_%7Dall.jsp?arnumber=5703374.

[67] M. H. Lee et al. "Ferroelectric negative capacitance hetero-tunnel field-effect-transistors with internal voltage amplification". In: *2013 IEEE International Electron Devices Meeting* (Dec. 2013), pp. 4.5.1–4.5.4. DOI: 10.1109/IEDM.2013.6724561. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6724561.

[68] M. H. Lee et al. "Prospects for ferroelectric HfZrOx FETs with experimentally CET=0.98nm, SSfor=42mV/dec, SSrev=28mV/dec, switch-off 0.2V, and hysteresis-free strategies". In: *2015 IEEE International Electron Devices Meeting (IEDM)*. IEEE, Dec. 2015, pp. 22.5.1–22.5.4. ISBN: 978-1-4673-9894-7. DOI: 10.1109/IEDM.2015.7409759. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7409759.

[69] Andrew K. Jonscher. "The physical origin of negative capacitance". In: *Journal of the Chemical Society, Faraday Transactions 2* 82.1 (1986), p. 75. ISSN: 0300-9238. DOI: 10.1039/f29868200075.

[70] John E. Hopcroft and Rajeev Motwani. *Introduction to Automata Theory, Languages, and Computation*. 3rd ed. Pearson, July 2006.

[71] Claude Elwood Shannon. "A symbolic analysis of relay and switching circuits". In: *Electrical Engineering* 57.12 (Dec. 1938), pp. 713–723. ISSN: 0095-9197. DOI: 10.1109/EE.1938.6431064.

[72] Yulong Leng. *Zhonghua Zihai*. 1994.

[73] Ministry of Education of the People's Republic of China. *List of Commonly Used Characters in Modern Chinese*. Jan. 1988.

[74] Ministry of Education of the People's Republic of China. *List of Frequently Used Characters in Modern Chinese*. Jan. 1988.

[75] John DeFrancis. *The Chinese Language: Fact and Fantasy*. 1st ed. University of Hawaii Press, Mar. 1986.

[76] Language Log. *Polysyllabic characters in Chinese writing*. Aug. 2011. URL: http://languagelog.ldc.upenn.edu/nll/?p=3330.

[77] E. Reischauer and J. Fairbank. *East Asia: The Great Tradition*. Houghton Mifflin, 1960.

[78] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. "Distributed Representations". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*. Cambridge: MIT Press, 1986. Chap. 3, pp. 77–109. ISBN: 0-262-68053-X.

[79] Tony Plate. "Holographic Reduced Representations: Convolutional Algebra for Compositional Distributed Representations". In: *Proceedings of the 12th International Joint Conference on Artificial Intelligence*. Ed. by John Mylopoulos, Ray Reiter, and Morgan Kaufmann. San Mateo, 1991, pp. 30–35. URL: http://www.ijcai.org/Proceedings/91-1/Papers/006.pdf.

[80] Gene Hecht. *Optics*. 5th ed. Pearson, Jan. 2016.

[81] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd ed. Wiley-Interscience, July 2006.

[82] H. S. Stone. "A Logic-in-Memory Computer". In: *IEEE Transactions on Computers* 19 (Jan. 1970), pp. 73–78. ISSN: 0018-9340. DOI: 10.1109/TC.1970.5008902. URL: doi.ieeecomputersociety.org/10.1109/TC.1970.5008902.

[83] Saugata Ghose et al. "The Processing-in-Memory Paradigm: Mechanisms to Enable Adoption". In: *Beyond-CMOS Technologies for Next Generation Computer Design*. Springer International Publishing, Jan. 2019, pp. 133–194. ISBN: 978-3-319-90384-2. DOI: 10.1007/978-3-319-90385-9_5.

[84] John Backus. "Can Programming Be Liberated from the Von Neumann Style?: A Functional Style and Its Algebra of Programs". In: *Commun. ACM* 21.8 (Aug. 1978), pp. 613–641. ISSN: 0001-0782. DOI: 10.1145/359576.359579. URL: http://doi.acm.org/10.1145/359576.359579.

[85] K. Hsieh et al. "Accelerating pointer chasing in 3D-stacked memory: Challenges, mechanisms, evaluation". In: *2016 IEEE 34th International Conference on Computer Design (ICCD)*. Oct. 2016, pp. 25–32. DOI: 10.1109/ICCD.2016.7753257.

[86] A. Boroumand et al. "LazyPIM: An Efficient Cache Coherence Mechanism for Processing-in-Memory". In: *IEEE Computer Architecture Letters* 16.1 (Jan. 2017), pp. 46–50. ISSN: 1556-6056. DOI: 10.1109/LCA.2016.2577557.

[87] Aditya Joshi, Johan T. Halseth, and Pentti Kanerva. "Language geometry using random indexing". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Ed. by Harald Atmanspacher et al. Vol. 10106 LNCS. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2017, pp. 265–274. ISBN: 9783319522883. DOI: 10.1007/978-3-319-52289-0_21. arXiv: arXiv:1412.7026v2. URL: http://link.springer.com/10.1007/978-3-319-52289-0%7B%5C_%7D21.

[88] David R Brillinger. "An Introduction to Polyspectra". In: *The Annals of Mathematical Statistics* 36.5 (Oct. 1965), pp. 1351–1374. ISSN: 0003-4851. DOI: 10.1214/aoms/1177699896. URL: http://www.jstor.org/stable/2238424%20http://projecteuclid.org/euclid.aoms/1177699896.

[89] M. Rosenblatt. "Cumulants and cumulant spectra". In: *Handbook of Statistics*. Vol. 3. 1983, pp. 369–382. ISBN: 9780444867261. DOI: 10.1016/S0169-7161(83)03019-9. URL: http://www.sciencedirect.com/science/article/pii/S0169716183030199%20http://linkinghub.elsevier.com/retrieve/pii/S0169716183030199.

[90] Gian-Carlo Rota and Jianhong Shen. "On the Combinatorics of Cumulants". In: *J. Comb. Theory Ser. A* 91.1 (July 2000), pp. 283–304. ISSN: 0097-3165. DOI: 10.1006/jcta.1999.3017. URL: http://dx.doi.org/10.1006/jcta.1999.3017.

[91] Roman Scoccimarro et al. "Nonlinear Evolution of the Bispectrum of Cosmological Perturbations". In: *The Astrophysical Journal* 496.2 (1998), p. 586. URL: http://stacks.iop.org/0004-637X/496/i=2/a=586.

[92] Ramakrishna Kakarala. "The Bispectrum as a Source of Phase-Sensitive Invariants for Fourier Descriptors: A Group-Theoretic Approach". In: *Journal of Mathematical Imaging and Vision* 44.3 (Nov. 2012), pp. 341–353. ISSN: 1573-7683. DOI: 10.1007/s10851-012-0330-6. URL: https://doi.org/10.1007/s10851-012-0330-6.

[93] N. Gillis and F. Glineur. "Low-Rank Matrix Approximation with Weights or Missing Data Is NP-Hard". In: *SIAM Journal on Matrix Analysis and Applications* 32.4 (2011), pp. 1149–1165. DOI: 10.1137/110820361. eprint: https://doi.org/10.1137/110820361. URL: https://doi.org/10.1137/110820361.

[94] Roman Vershynin. 1st ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Sept. 2018.

[95] Robert W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer, Sept. 2010.

[96] Gian-Carlo Rota, Jianhong Shen, and Brian D. Taylor. "All polynomials of binomial type are represented by Abel polynomials". en. In: *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze* Ser. 4, 25.3-4 (1997), pp. 731–738. URL: http://http://www.numdam.org/item/ASNSP_1997_4_25_3-4_731_0.

[97] E. Di Nardo, P. Petrullo, and D. Senato. "Cumulants and convolutions via Abel polynomials". In: *European Journal of Combinatorics* 31.7 (2010), pp. 1792–1804. ISSN: 0195-6698. DOI: https://doi.org/10.1016/j.ejc.2010.03.002. URL: http://www.sciencedirect.com/science/article/pii/S0195669810000302.

[98] Audio Engineering Society. *AES recommended practice for professional digital audio - Preferred sampling frequencies for applications employing pulse-code modulation*. 2013.

[99] C. Eckart and G. Young. "The approximation of one matrix by another of lower rank". In: *Psychometrika* 1.3 (1936), pp. 211–218. DOI: 10.1007/BF02288367.

[100] L. Mirsky. "Symmetric Gauge Functions and Unitarily Invariant Norms". In: *The Quarterly Journal of Mathematics* 11.1 (1960), pp. 50–59. DOI: 10.1093/qmath/11.1.50. eprint: /oup/backfile/content_public/journal/qjmath/11/1/10.1093_qmath_11.1.50/3/11-1-50.pdf. URL: http://dx.doi.org/10.1093/qmath/11.1.50.

[101] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[102] G. Tzanetakis and P. Cook. "Musical genre classification of audio signals". In: *IEEE Transactions on Speech and Audio Processing* 10.5 (July 2002), pp. 293–302. ISSN: 1063-6676. DOI: 10.1109/TSA.2002.800560.

[103] Jean-Julien Aucouturier. "Representing Musical Genre: A State of the Art". In: *Journal of New Music Research* 32 (Mar. 2003), pp. 83–93. DOI: 10.1076/jnmr.32.1.83.16801.

[104] Elias Pampalk, Simon Dixon, and Gerhard Widmer. "On the Evaluation of Perceptual Similarity Measures for Music". In: *In Proceedings of the International Conference on Digital Audio Effects (DAFx-03)*. 2003, pp. 7–12.

[105] Fabian Mörchen et al. “Databionic Visualization of Music Collections According to Perceptual Distance”. In: *ISMIR*. 2005.