

On loss landscapes of error-correcting codes: symmetries, linear codes and stationarity

R. Shekhar¹, A. Mulgund¹, N. Devroye¹, Gy. Turán^{1,2}, M. Žefran¹

¹University of Illinois Chicago, Chicago, IL, USA

²HUN-REN–SZTE Research Group on AI, University of Szeged, Hungary

Email: rshekh3@uic.edu, mulgund2@uic.edu, devroye@uic.edu, gyt@uic.edu, mzefran@uic.edu

Abstract—Motivated by machine learning approaches to error-correcting codes, we consider finding good error-correcting codes as an optimization problem over a space of codebooks, where the loss function depends on the channel, the decoding method and the error measure. General results are proven assuming symmetry and mild differentiability properties of the loss function. For binary linear codes a stationarity result is given in terms of power reallocations of the code. For transitive binary linear codes such a result is given for the code itself.

I. INTRODUCTION

Traditionally, error-correcting codes have been constructed using algebraic and combinatorial methods. In the last decade, machine learning made it possible to jointly learn an encoder and a decoder using the channel autoencoder framework [1]. The encoder and the decoder are both represented as neural networks, and a noise layer in between represents the channel.

Finding a good code using this approach amounts to optimization over the set of codes representable by a network architecture with some choice of network parameters. The optimization objective is to minimize a *loss function*, which is some *error measure* of the code, like BLER or its surrogates.

Understanding the structure of the loss landscapes of learning problems for neural networks is an active area of deep learning research [2, Ch. 12]. This involves determining stationary points, local minima, their flatness, and other features that determine the fixed points and training trajectories of gradient-based learning algorithms.

For error-correcting codes one can consider a loss landscape over a more general structure: the set of *all codes* with a given input size and code length. Codes representable by a given network form a subset, and so local minima of the set of all codes remain local minima for the network loss landscape. Thus, understanding the loss landscapes over the set of all codes is a first step towards understanding the loss landscapes over neural networks implementing codes.

The general setup to study the loss landscape is formulated by representing an $[n, k]$ code of length n and $M = 2^k$ messages as a real-valued matrix $\mathbf{X} \in \mathbb{R}^{n \times M}$ called the *codebook*. Real-valued codes jointly code and modulate the M messages, and this is the natural setting in which deep learning is applied to code design [3], [4]. Taking the power constraint into consideration as well, the set of codes is turned into a manifold; for average codeword power constraint the

manifold is a Frobenius sphere. The channel, the decoding method, and the error measure are assumed to be fixed. In this paper, we will develop a *general framework* for studying stationarity of codes built on symmetries of these fixed pieces. However, for concreteness, a basic case to consider is average power constraint, AWGN channel, maximum likelihood (ML) decoding and BLER error measure. With this power constraint, optimization is over the $Mn - 1$ dimensional Frobenius sphere. In this case the loss is a function of the codebook only.

An important class of codes is binary linear codes. In what follows linear codes will always be assumed to be binary. Experimental results suggest that linear codes may be local minima [5]. In this paper we consider some experimental and theoretical aspects of this question.

The discussion of theoretical results begins with introducing the formal framework for error-correcting code optimization in Section III. Note that this is not meant as a realistic model for practical optimization, but rather as a conceptual model underlying machine learning approaches.

We give experimental results in Section IV-A suggesting that Reed-Muller (RM) codes are stationary. These experiments reinforce the observations in [5]. However, it turns out that the local minimality question is quite subtle. In Section IV-B we present simple examples showing that some linear codes are not even stationary.

We present two theoretical results on stationarity for linear codes under mild differentiability assumptions:

- Every nondegenerate linear code has a distinguished power allocation which is stationary, if differentiable. It is always differentiable for AWGN-ML-BLER (Section V).
- Every nondegenerate linear code with a transitive automorphism group is stationary (Section VI). RM codes and cyclic codes are special cases.

The proofs are based on studying the implications of *symmetries of codes*. As far as we know, the approach of this paper provides a new lens for the study of real-valued error-correcting codes.

II. RELATED WORK

The study of error-correcting code symmetries is a classical topic in coding theory [6]. A general result on capacity-achieving codes based on double transitivity is given in [7].

Global optimality of spherical codes [8] has been studied extensively as a constrained optimization problem over point configurations on the sphere. Pairwise energies of the form $E_h(\mathbf{X}) = \sum_{1 \leq i < j \leq M} h(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$, where h is a *completely monotone potential*, are considered in [9]. It is proven that sharp configurations minimize E_h for such potentials. This applies, in particular, to *cross polytopes*, and provides a global optimality result for $\text{RM}(m, 1)$ codes, as their codebook forms a cross polytope [10]. Our focus is different as we do not consider global optimality, but instead we study aspects relevant for machine learning, such as stationarity. Tokarchuk et al. [11] give an overview of recent applications and machine learning aspects of the classical problem.

The question “do symmetric problems have symmetric solutions?” [12] has been considered in many contexts. The general answer is negative, but positive answers hold under reasonable conditions, as in [12] and Palais’ principle of symmetric criticality [13]. This turns out to be the case in this paper as well. The closest analog of our framework is the equivariant-dynamics perspective of Golubitsky and Stewart [14], where the invariance of fixed-point subspaces under equivariant dynamics plays the same structural role as Lemma V.2. Related symmetry methods for differential equations are developed in [15]–[17]. Closely related fixed-point arguments have been used to study deep-learning loss landscapes [18], [19], where neuron-permutation and equivariance-induced symmetries similarly constrain stationary point properties; [20] treats different symmetries of linear autoencoders and gives a theoretical description of all stationary points.

This paper continues our previous work on understanding deep-learned error-correcting codes [5], [21]–[27].

III. SETUP: CODEBOOK, LOSSES, SYMMETRIES, AND POWER CONSTRAINTS

In this section we formulate the general framework for studying loss landscapes of error-correcting codes.

A. Real-valued codebooks and power constraints

We represent a real-valued codebook of blocklength n and size M as $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_M] \in \mathbb{R}^{n \times M}$, where $\mathbf{x}_i \in \mathbb{R}^n$. A codebook is *nondegenerate* if all codewords are distinct.

A binary block code $\{0, 1\}^k \rightarrow \{0, 1\}^n$ is embedded into this real space using the standard BPSK map $0 \mapsto +1$ and $1 \mapsto -1$, giving codebooks in $\{\pm 1\}^{n \times M} \subseteq \mathbb{R}^{n \times M}$ with $M = 2^k$. When the binary code is linear, its BPSK image must contain the all-ones codeword, and w.l.o.g. we index this codeword as $\mathbf{x}_1 = \mathbf{1}$. Unless otherwise specified, binary linear codes are assumed to be embedded in real space via BPSK, and we refer to the embedded code as a linear code as well.

We impose the *average codeword power* constraint

$$\frac{1}{M} \sum_{i=1}^M \|\mathbf{x}_i\|_2^2 = n, \quad \text{equivalently} \quad \|\mathbf{X}\|_F^2 = nM \quad (1)$$

where $\|\mathbf{X}\|_F^2 = \langle \mathbf{X}, \mathbf{X} \rangle_F = \sum_{i=1}^M \|\mathbf{x}_i\|_2^2$ is the squared Frobenius norm of $\mathbf{X}$ induced by the Frobenius inner product

$\langle A, B \rangle_F := \text{Tr}(A^T B)$. This allows power to be redistributed across codewords, and hence is more flexible than the per-codeword constraint ($\forall i \|\mathbf{x}_i\|_2^2 = n$) of spherical codes.

B. Losses and symmetries

For a fixed noisy channel, a codeword $\mathbf{x}_j \in \mathbb{R}^n$ is communicated by n transmissions over the noisy channel. We assume that the performance of the codebook $\mathbf{X}$ is determined by some *loss function* $f : \mathbb{R}^{n \times M} \rightarrow \mathbb{R}$, which is assumed to be continuous everywhere and differentiable at nondegenerate points. We consider situations where performance of the codebook $\mathbf{X}$ is invariant under reordering of codewords, reordering of the n coordinate-wise transmissions as well as consistent sign flips of all codewords. Thus f is invariant under the following symmetries:

$$f(\mathbf{P}\mathbf{X}) = f(\mathbf{X}) \quad \text{for every } \mathbf{P} \in S_n \quad (2)$$

$$f(\mathbf{D}\mathbf{X}) = f(\mathbf{X}) \quad \text{for every } \mathbf{D} \in \text{diag}(\{\pm 1\}^n) \quad (3)$$

$$f(\mathbf{X}\mathbf{\Pi}) = f(\mathbf{X}) \quad \text{for every } \mathbf{\Pi} \in S_M \quad (4)$$

where S_n, S_M are the permutation matrix groups on n, M coordinates resp., and $\text{diag}(\{\pm 1\}^n)$ is the set of diagonal matrices with diagonal entries in $\{\pm 1\}$. We call the group generated by $S_n \cup \text{diag}(\{\pm 1\}^n)$ the *signed permutation group* and denote it $S_n^\pm$. $S_n^\pm$ is the symmetry group of the n -dimensional hypercube and cross-polytope [28, p. 122].

C. The error-correcting code optimization problem

Now we can formulate the general problem considered in this paper. Let $f : \mathbb{R}^{n \times M} \rightarrow \mathbb{R}$ be a continuous function satisfying symmetry conditions (2)–(4). The *error-correcting code optimization* problem is then:

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times M}} f(\mathbf{X}) \quad \text{s.t.} \quad \|\mathbf{X}\|_F^2 = nM. \quad (5)$$

We give a brief description of the geometric view of the optimization problem.

Codebooks $\mathbf{X}$ that satisfy (1) are called *feasible*, and they form the *feasible set* $\mathcal{M}_{\text{avg}} := \{\mathbf{X} \in \mathbb{R}^{n \times M} : \|\mathbf{X}\|_F^2 = nM\}$, which is the *Frobenius sphere* of radius $\sqrt{nM}$.

Since $\mathcal{M}_{\text{avg}}$ is a sphere, the subspace $\mathbf{X}^\perp$ orthogonal to $\mathbf{X}$ w.r.t. $\langle \cdot, \cdot \rangle_F$ forms the vector space tangent to $\mathcal{M}_{\text{avg}}$ at $\mathbf{X}$. $\text{T}_{\mathbf{X}}\mathcal{M}_{\text{avg}} = \mathbf{X}^\perp := \{\Delta \in \mathbb{R}^{n \times M} : \langle \Delta, \mathbf{X} \rangle_F = 0\}$.

For a feasible $\mathbf{X}$, let $G = \nabla f(\mathbf{X}) \in \mathbb{R}^{n \times M}$ be the gradient of f at $\mathbf{X}$. Then the *tangent gradient* at $\mathbf{X}$, $\nabla_{\text{tan}} f(\mathbf{X}) \in \text{T}_{\mathbf{X}}\mathcal{M}_{\text{avg}}$, can be obtained by removing the component along $\mathbf{X}$, i.e. $\nabla_{\text{tan}} f(\mathbf{X}) = G - \frac{\langle G, \mathbf{X} \rangle_F}{\|\mathbf{X}\|_F^2} \mathbf{X} = G - \frac{\langle G, \mathbf{X} \rangle_F}{nM} \mathbf{X}$.

A feasible codebook $\mathbf{X} \in \mathcal{M}_{\text{avg}}$ is a *stationary* point of the constrained loss f in (5) if $\nabla_{\text{tan}} f(\mathbf{X}) = 0$. This occurs precisely when $\nabla f(\mathbf{X}) = \alpha \mathbf{X}$ for some $\alpha \in \mathbb{R}$.

At $\mathbf{X} \in \mathcal{M}_{\text{avg}}$, a *feasible ϵ -perturbation* $\mathbf{X}_\epsilon$, $\epsilon \in \mathbb{R}$, along a given tangent vector $\Delta \in \text{T}_{\mathbf{X}}\mathcal{M}_{\text{avg}}$ can be found as $\mathbf{X}_\epsilon := \mathbf{R}_{\mathbf{X}}(\epsilon \Delta)$, where $\mathbf{R}_{\mathbf{X}} : \text{T}_{\mathbf{X}}\mathcal{M}_{\text{avg}} \rightarrow \mathcal{M}_{\text{avg}}$ $\mathbf{R}_{\mathbf{X}}(\Delta) := \sqrt{nM} (\mathbf{X} + \Delta) / (\|\mathbf{X} + \Delta\|_F)$ is the *normalization retraction*, which makes local tangential perturbations and rescales them back onto the feasible set.

These feasible perturbations can be used to optimize (5) via Riemannian gradient descent (RGD) [29, Ch. 4], whose iterates $\{\mathbf{X}^{(t)}\}_{t \geq 0}$ are initialized at some $\mathbf{X}^{(0)} \in \mathcal{M}_{\text{avg}}$ and are then given by $\mathbf{X}^{(t+1)} = \mathbf{R}_{\mathbf{X}^{(t)}}(\epsilon^{(t)} \Delta^{(t)})$, where $\Delta^{(t)} = -\nabla_{\text{tan}} f(\mathbf{X}^{(t)})$ with $\epsilon^{(t)} > 0$.

Standing conventions. Hereafter, unless stated otherwise:

- f satisfies the symmetries (2)-(4),
- f is differentiable at every nondegenerate codebook, so derivatives can always be taken at those points,
- stationarity is with respect to the constrained problem (5).

D. Examples

In order to specify an instance of the optimization problem, the loss f induced by the channel noise, the decoding rule, and the chosen error measure need to be specified. The invariance assumptions (2)-(4) are satisfied, for example, when the channel is an additive noise channel with i.i.d. noise with even density, the decoder is ML, the messages are equiprobable, and the error measure is relabeling invariant.

A standard example covered by the above assumptions is AWGN-ML-BLER with uniform prior on messages.

Definition III.1 (AWGN-ML-BLER).

Draw codeword index J uniformly from $1, \dots, M$, transmit $Y = \mathbf{x}_J + Z$, where $Z \sim \mathcal{N}(0, \sigma^2 I_n)$, and decode $\hat{J}(Y) = \arg \min_j \|Y - \mathbf{x}_j\|_2$, with ties broken arbitrarily. Then,

$$\text{AWGN-ML-BLER}(\mathbf{X}) := \Pr\{\hat{J}(Y) \neq J\}.$$

Proposition III.2.

AWGN-ML-BLER is continuous on $\mathbb{R}^{n \times M}$ and is differentiable at every nondegenerate codebook.

Proof Sketch. Let ϕ_σ be the $\mathcal{N}(0, \sigma^2 I_n)$ density. The correct-decoding probability $P_c(\mathbf{X})$ is

$$P_c(\mathbf{X}) = 1 - \text{BLER}(\mathbf{X}) = \frac{1}{M} \int_{\mathbb{R}^n} \max_{j \in [M]} \phi_\sigma(y - \mathbf{x}_j) dy. \quad (6)$$

This formula is independent of tie-breaking and is continuous in $\mathbf{X}$. At a nondegenerate codebook, the maximizing index is unique except on a measure-zero set. Differentiating under the Gaussian integral gives differentiability. $\square$

Another example, used in our experiments, is BLER_{UB} , a surrogate for BLER obtained by applying the union bound as described below. It also satisfies conditions (2)-(4) and is differentiable at nondegenerate codebooks.

If $\mathbf{x}_i$ is transmitted and $\mathbf{x}_j$ is a competing codeword, the pairwise error probability under ML nearest-neighbor decoding is $Q\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma}\right)$, where $Q(\cdot)$ is the standard Gaussian tail function. Averaging over transmitted messages, we get

$$\text{BLER}_{\text{UB}}(\mathbf{X}) = \frac{2}{M} \sum_{1 \leq i < j \leq M} Q\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{2\sigma}\right) \geq \text{BLER}(\mathbf{X}).$$

This function depends only on pairwise distances between codewords, and can be evaluated and differentiated exactly, without a Monte Carlo simulation over the noisy channel.

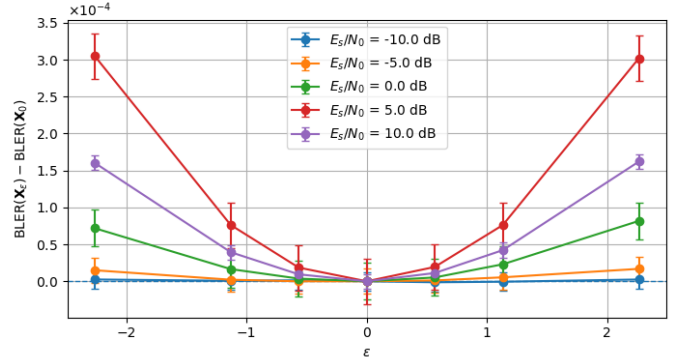


Fig. 1: Estimated ML-BLER variation along random feasible perturbations of the RM(5,1) codebook over the AWGN channel using ML decoding. The plotted quantity is $\text{BLER}(\mathbf{X}_\epsilon) - \text{BLER}(\mathbf{X}_0)$, with the average codeword power constraint enforced by global normalization of the codebook. Each SNR uses a separately sampled random tangent direction, so the curves represent independent local probes.

IV. EXPERIMENTS

We now give two numerical experiments illustrating the stationarity behavior of linear codes. The first gives supporting evidence from random tangent-direction probes. Theoretical justification is given later via Theorem VI.1. The second shows that not all linear codes are stationary.

A. Reed-Muller codes appear stationary

Reed-Muller codes $\text{RM}(m, r)$ are among the most important families of error-correcting codes [30]. We numerically probe the local behavior of the RM(5,1) codebook ($n = 32$ and $M = 2^6 = 64$) for BLER as the loss function, at various SNR values.

Let $\mathbf{X}_0$ denote the RM(5,1) codebook. Since we cannot directly compute the true tangent gradient $\nabla_{\text{tan}} \text{BLER}(\mathbf{X}_0)$, we estimate the local change in BLER around $\mathbf{X}_0$ along random perturbation directions by Monte Carlo simulation. Thus for each SNR value, we sample a random perturbation direction $\Delta = [\delta_1, \dots, \delta_M]$ in the tangent space at $\mathbf{X}_0$. For several $\epsilon \in [-2, 2]$, we compute $\mathbf{X}_\epsilon$. For each value of ϵ , we estimate the local change in BLER by Monte Carlo simulation, *i.e.* messages are sampled uniformly, AWGN noise is added, and decoding is performed using the ML decoder. Figure 1 plots the variations $\text{BLER}(\mathbf{X}_\epsilon) - \text{BLER}(\mathbf{X}_0)$, at different SNR values, so the curves show the local BLER change relative to the original RM(5,1) codebook.

The observed bowl-shaped profiles suggest that, along the sampled average-power-feasible perturbations, moving away from RM(5,1) tends to increase BLER. Similar observations were made for other small blocklength RM($m, 1$) codes.

B. Not all linear codes are stationary points

We enumerate a small family of inequivalent binary linear codes. Each code $\mathcal{C}^{(i)} \subseteq \mathbb{F}_2^4$ is embedded into a real codebook

$\mathbf{X}^{(i)} \in \mathbb{R}^{4 \times 4}$ using the standard BPSK map, with one codeword per column. We use BLER_{UB} from Section III-D as the loss function. We evaluate stationarity of codebooks $\mathbf{X}^{(i)}$ by computing the tangent gradient and testing whether its norm is zero, *i.e.* whether $\|\nabla_{\text{tan}} \text{BLER}_{\text{UB}}(\mathbf{X}^{(i)})\|_F \approx 0$.

The results are shown in Table I. Among the tested inequivalent linear $[4, 2]$ codes $\mathcal{C}^{(1)}, \dots, \mathcal{C}^{(6)}$, only $\mathcal{C}^{(1)}$ has numerically vanishing tangent gradient at the tested SNR of 0 dB. The table also shows that stationarity does not imply the best BLER_{UB} performance: the stationary code $\mathcal{C}^{(1)}$ is dominated by the nonstationary code $\mathcal{C}^{(2)}$.

Since the other linear codes ($\neq \mathcal{C}^{(1)}$) have nonzero tangent gradients, their codebooks admit feasible first-order descent directions. We run RGD initialized at each codebook $\mathbf{X}^{(i)}$ until convergence. The resulting optimized codebooks $\mathbf{X}_*^{(i)}$ have vanishing tangent gradients, so they all appear stationary.

For all i , $\mathbf{X}_*^{(i)} \approx \text{diag}(\boldsymbol{\mu}_*^{(i)})\mathbf{X}^{(i)}$ for some rescaling vector $\boldsymbol{\mu}_*^{(i)}$, see Table I row 7, *i.e.* the resulting optimized codebooks $\mathbf{X}_*^{(i)}$ are coordinate-wise rescalings of the original codebooks $\mathbf{X}^{(i)}$. In the reported runs, the rescaling vectors satisfy $\|\boldsymbol{\mu}_*^{(i)}\|^2 = n$, so the resulting codebooks satisfy the average codeword power constraint. We call such a $\boldsymbol{\mu}$ a *power reallocation*, since it redistributes the power budget across the n coordinates. This phenomenon is a consequence of linearity and the symmetries of the loss, and related theoretical aspects are discussed in Section V. It is worth noting that for $\mathcal{C}^{(3)}, \mathcal{C}^{(4)}$, and $\mathcal{C}^{(6)}$, RGD assigns zero power to precisely those coordinates which are constant across all codewords, and so they carry no information. The algorithm thus “discovered” puncturing.

V. STATIONARITY OF LINEAR CODES WITH POWER REALLOCATION

Section IV-B showed empirically that BPSK-mapped linear codes need not be stationary, yet RGD from each tested code converged to a power reallocation of the same codebook (Row 7 of Table I). We now show that this is the case in general. For a power allocation $\boldsymbol{\mu} \in \mathbb{R}^n$, let $\mathbf{X}_{\boldsymbol{\mu}} := \text{diag}(\boldsymbol{\mu})\mathbf{X}$, and let $\mathcal{S}_n := \{\boldsymbol{\mu} : \|\boldsymbol{\mu}\|_2^2 = n\}$. Denote the feasible power reallocations of $\mathbf{X}$ by $\mathcal{F}_{\mathbf{X}} = \{\mathbf{X}_{\boldsymbol{\mu}} : \boldsymbol{\mu} \in \mathcal{S}_n\}$. Since $\mathbf{1} \in \mathcal{S}_n$, the original code $\mathbf{X} = \mathbf{X}_1$ is in the feasible set $\mathcal{F}_{\mathbf{X}}$. Since $\mathcal{S}_n$ is compact and $\boldsymbol{\mu} \mapsto f(\mathbf{X}_{\boldsymbol{\mu}})$ is continuous, a minimizer exists; denote one by $\boldsymbol{\mu}_*$ and its reallocation by $\mathbf{X}_{\boldsymbol{\mu}_*}$.

Theorem V.1.

Let $\mathbf{X} \in \mathbb{R}^{n \times M}$ be a nondegenerate linear code. If f is differentiable at $\mathbf{X}_{\boldsymbol{\mu}_*}$, then $\mathbf{X}_{\boldsymbol{\mu}_*}$ is a stationary point.

In other words, $\mathbf{X}_{\boldsymbol{\mu}_*}$ is stationary for the feasible power allocations $\mathcal{F}_{\mathbf{X}}$ if and only if it is stationary over the full feasible set $\mathcal{M}_{\text{avg}}$.

To prove Theorem V.1, we need some new concepts. An *automorphism* of a code $\mathbf{X}$ is some $\mathbf{P} \in S_n$ such that $\mathbf{P}\mathbf{X}\mathbf{\Pi}^{-1} = \mathbf{X}$ for some $\mathbf{\Pi} \in S_M$ [6, Ch. 8.5]. The group of code automorphisms is denoted $\text{Aut}(\mathbf{X})$. We generalize

this notion: a *bi-automorphism* is some $(\mathbf{R}, \mathbf{\Pi}) \in S_n^{\pm} \times S_M$ such that $\mathbf{R}\mathbf{X}\mathbf{\Pi}^{-1} = \mathbf{X}$ (recall $S_n^{\pm}$ is the group of signed permutation matrices). The bi-automorphisms of $\mathbf{X}$ form a group, which we denote $\text{Aut}_{\text{bi}}(\mathbf{X})$.

The fixed-point subspace of $\text{Aut}_{\text{bi}}(\mathbf{X})$ is

$$\text{Fix}(\text{Aut}_{\text{bi}}(\mathbf{X})) := \{\mathbf{A} \in \mathbb{R}^{n \times M} : \mathbf{R}\mathbf{A}\mathbf{\Pi}^{-1} = \mathbf{A} \text{ for every } (\mathbf{R}, \mathbf{\Pi}) \in \text{Aut}_{\text{bi}}(\mathbf{X})\}.$$

More generally, for any group Γ acting by linear isomorphism on $\mathbb{R}^{n \times M}$, $\text{Fix}(\Gamma)$ consists of those $\mathbf{A} \in \mathbb{R}^{n \times M}$ fixed by all $g \in \Gamma$. It is always a vector subspace of $\mathbb{R}^{n \times M}$. Note that if $\Gamma_2 \subseteq \Gamma_1$, we have the reversed inclusion $\text{Fix}(\Gamma_1) \subseteq \text{Fix}(\Gamma_2)$.

For a linear code $\mathbf{X}$, adding $(\oplus)$ a codeword $\mathbf{x}_k$ to all codewords gives the same codewords up to permutation. In BPSK this means elementwise multiplication $(\odot)$ of $\mathbf{x}_k$ with every codeword, *i.e.* $[\mathbf{x}_k \odot \mathbf{x}_1, \dots, \mathbf{x}_k \odot \mathbf{x}_M]$. Note that $\mathbf{x}_k \odot \mathbf{x} = \text{diag}(\mathbf{x}_k)\mathbf{x}$ for any vector $\mathbf{x}$. Define $\mathbf{D}_k := \text{diag}(\mathbf{x}_k)$. Note $\mathbf{D}_k \in S_n^{\pm}$ since $\mathbf{x}_k \in \{\pm 1\}^n$. Thus $\mathbf{D}_k\mathbf{X}$ is a permutation of the columns of $\mathbf{X}$, and so there exists a permutation $\mathbf{\Pi}_k \in S_M$ such that $\mathbf{D}_k\mathbf{X} = \mathbf{X}\mathbf{\Pi}_k$, equivalently $\mathbf{D}_k\mathbf{X}\mathbf{\Pi}_k^{-1} = \mathbf{X}$. Linearity of $\mathbf{X}$ ensures $\{(\mathbf{D}_k, \mathbf{\Pi}_k)\}_{k=1}^M$ is closed under composition. Thus we get a subgroup of bi-automorphisms of $\mathbf{X}$:

$$\text{Aut}_{\text{LIN}}(\mathbf{X}) := \{(\mathbf{D}_k, \mathbf{\Pi}_k) : k = 1, \dots, M\} \subseteq \text{Aut}_{\text{bi}}(\mathbf{X}).$$

We now show how invariance of the loss function constrains the gradient at a codebook $\mathbf{X}$.

Lemma V.2.

If f is differentiable at $\mathbf{X}$, then $\nabla f(\mathbf{X}) \in \text{Fix}(\text{Aut}_{\text{bi}}(\mathbf{X}))$.

Proof. Let $(\mathbf{R}, \mathbf{\Pi}) \in \text{Aut}_{\text{bi}}(\mathbf{X})$. By invariances (2)-(4) of f , $f(\mathbf{Y}) = f(\mathbf{R}\mathbf{Y}\mathbf{\Pi}^{-1}) \forall \mathbf{Y} \in \mathbb{R}^{n \times M}$. Differentiating both sides *w.r.t.* $\mathbf{Y}$ at $\mathbf{Y} = \mathbf{X}$ (via chain rule on the right) gives $\nabla f(\mathbf{X}) = \mathbf{R}^T(\nabla f(\mathbf{R}\mathbf{X}\mathbf{\Pi}^{-1}))(\mathbf{\Pi}^{-1})^T$. By assumption/definition, $\mathbf{R}, \mathbf{\Pi}$ are orthogonal, and $\mathbf{R}\mathbf{X}\mathbf{\Pi}^{-1} = \mathbf{X}$. Rearranging terms we get $\nabla f(\mathbf{X}) = \mathbf{R}\nabla f(\mathbf{X})\mathbf{\Pi}^{-1}$. $\square$

Lemma V.3.

For linear $\mathbf{X}$, $\text{Fix}(\text{Aut}_{\text{LIN}}(\mathbf{X})) = \{\mathbf{X}_{\boldsymbol{\mu}} : \boldsymbol{\mu} \in \mathbb{R}^n\}$.

Proof. We prove this by showing inclusion in both directions.

$\subseteq$: Note that $\mathbf{D}_k$ swaps $\mathbf{x}_1 = \mathbf{1}$ with $\mathbf{x}_k$, as $\mathbf{D}_k[\mathbf{x}_1 \dots \mathbf{x}_k \dots] = [\underbrace{\mathbf{D}_k\mathbf{x}_1}_{\mathbf{x}_k \odot \mathbf{x}_1} \dots \underbrace{\mathbf{D}_k\mathbf{x}_k}_{\mathbf{x}_k \odot \mathbf{x}_k} \dots] = [\mathbf{x}_k \dots \mathbf{x}_1 \dots]$.

Thus $\mathbf{\Pi}_k^{-1}$ must also do so, as $(\mathbf{D}_k, \mathbf{\Pi}_k) \in \text{Aut}_{\text{LIN}}(\mathbf{X})$. Let $\mathbf{Y} \in \text{Fix}(\text{Aut}_{\text{LIN}}(\mathbf{X}))$, then we have $\mathbf{D}_k\mathbf{Y}\mathbf{\Pi}_k^{-1} = \mathbf{Y}$, thus $\mathbf{y}_k = (\mathbf{D}_k\mathbf{Y}\mathbf{\Pi}_k^{-1})_{:,k} = (\mathbf{D}_k\mathbf{Y})_{:,1} = \mathbf{D}_k\mathbf{y}_1 = \text{diag}(\mathbf{y}_1)\mathbf{x}_k$. Stacking all columns gives $\mathbf{Y} = \text{diag}(\mathbf{y}_1)\mathbf{X} \in \{\mathbf{X}_{\boldsymbol{\mu}} : \boldsymbol{\mu} \in \mathbb{R}^n\}$.

$\supseteq$: Since diagonal matrices commute, $\mathbf{D}_k(\mathbf{X}_{\boldsymbol{\mu}})\mathbf{\Pi}_k^{-1} = \text{diag}(\boldsymbol{\mu})(\mathbf{D}_k\mathbf{X}\mathbf{\Pi}_k^{-1}) = \mathbf{X}_{\boldsymbol{\mu}}$, so for $\boldsymbol{\mu} \in \mathbb{R}^n$, $\mathbf{X}_{\boldsymbol{\mu}} \in \text{Fix}(\text{Aut}_{\text{LIN}}(\mathbf{X}))$. $\square$

Corollary V.4.

For a linear code $\mathbf{X}$, $\text{Fix}(\text{Aut}_{\text{bi}}(\mathbf{X})) \subseteq \{\mathbf{X}_{\boldsymbol{\mu}} : \boldsymbol{\mu} \in \mathbb{R}^n\}$.

Row #	Quantity	$\mathcal{C}^{(1)}$	$\mathcal{C}^{(2)}$	$\mathcal{C}^{(3)}$	$\mathcal{C}^{(4)}$	$\mathcal{C}^{(5)}$	$\mathcal{C}^{(6)}$
1	Code	$\begin{Bmatrix} 0000, 0011 \\ 1100, 1111 \end{Bmatrix}$	$\begin{Bmatrix} 0000, 0011 \\ 1101, 1110 \end{Bmatrix}$	$\begin{Bmatrix} 0000, 0011 \\ 0101, 0110 \end{Bmatrix}$	$\begin{Bmatrix} 0000, 0001 \\ 0010, 0011 \end{Bmatrix}$	$\begin{Bmatrix} 0000, 0001 \\ 1110, 1111 \end{Bmatrix}$	$\begin{Bmatrix} 0000, 0001 \\ 0110, 0111 \end{Bmatrix}$
2	$A(x)$	$1 + 2x^2 + x^4$	$1 + x^2 + 2x^3$	$1 + 3x^2$	$1 + 2x + x^2$	$1 + x + x^3 + x^4$	$1 + x + x^2 + x^3$
3	$d_{\min}$	2	2	2	1	1	1
4	Stationary?	Y	N	N	N	N	N
5	BLER _{UB}	4.784×10^{-2}	3.706×10^{-2}	6.825×10^{-2}	1.800×10^{-1}	8.814×10^{-2}	1.086×10^{-1}
6	$\ \nabla_{\tan}\ $	0.00×10^0	1.89×10^{-2}	4.68×10^{-2}	1.31×10^{-1}	8.29×10^{-2}	8.18×10^{-2}
7	$(\mu_*^{(i)})^T$	$[1, 1, 1, 1]$	$\sqrt{\frac{2}{3}}[1, 1, \sqrt{2}, \sqrt{2}]$	$\sqrt{\frac{4}{3}}[0, 1, 1, 1]$	$\sqrt{2}[0, 0, 1, 1]$	$\sqrt{\frac{2}{3}}[1, 1, 1, \sqrt{3}]$	$[0, 1, 1, \sqrt{2}]$

TABLE I: Summary of stationarity and surrogate BLER results for inequivalent $[4, 2]$ linear codes over the AWGN channel at 0 dB. **Columns** correspond to one code $\mathcal{C}^{(i)}$. **Row 1** lists the codewords in $\mathcal{C}^{(i)} \subseteq \mathbb{F}_2^4$; colored coordinates are constant across all codewords. **Row 2** shows the Hamming weight enumerator $A(x)$. **Row 3** shows the minimum Hamming distance $d_{\min}$. **Row 4** shows whether the codebook $\mathbf{X}^{(i)}$ corresponding to $\mathcal{C}^{(i)}$ is numerically stationary. **Row 5** shows the corresponding surrogate values BLER_{UB}. Only $\mathcal{C}^{(1)}$ is numerically stationary, while the nonstationary code $\mathcal{C}^{(2)}$ has lower BLER_{UB}. Thus stationarity and BLER_{UB} performance do not necessarily coincide. **Row 6** shows the tangent-gradient norms at the real-valued codebook for $\mathcal{C}^{(i)}$ at 0 dB. **Row 7** shows the coordinate-wise rescaling $\mu_*^{(i)}$ s.t. $\mathbf{X}_*^{(i)} = \text{diag}(\mu_*^{(i)})\mathbf{X}^{(i)}$, where $\mathbf{X}_*^{(i)}$ is the result of RGD optimization initialized at $\mathbf{X}^{(i)}$. Colored coordinates are reduced to zero and coincide with those in Row 1.

Proof of Theorem V.1. Let $L(\mu) := f(\mathbf{X}_\mu)$ on $\mathcal{S}_n$. The sphere $\mathcal{S}_n$ is compact, so continuity gives a minimizer μ_* . Assuming f is differentiable at $\mathbf{X}_{\mu_*}$, the reduced function L is differentiable at μ_* . The first-order optimality condition on the sphere gives some $\lambda \in \mathbb{R}$ with $\nabla L(\mu_*) + 2\lambda\mu_* = 0$. By Lemma V.3, $\mathbf{X}_{\mu_*} \in \text{Fix}(\text{Aut}_{\text{LIN}}(\mathbf{X}))$, so $\text{Aut}_{\text{LIN}}(\mathbf{X}) \subseteq \text{Aut}_{\text{bi}}(\mathbf{X}_{\mu_*})$. Lemmas V.2 and V.3 then give $\nabla f(\mathbf{X}_{\mu_*}) = \text{diag}(\mathbf{d})\mathbf{X}$ for some $\mathbf{d} \in \mathbb{R}^n$. Observe that $\partial_{\mu_i} L(\mu_*) = \langle \partial_{\mathbf{X}_{i,:}} f(\mathbf{X}_{\mu_*}), \mathbf{X}_{i,:} \rangle \quad \forall i \in [n]$. Then, $\partial_{\mu_i} L(\mu_*) = \mathbf{d}_i \|\mathbf{X}_{i,:}\|_2^2 = M\mathbf{d}_i$, since $\mathbf{X} \in \{\pm 1\}^{n \times M}$. Thus $-2\lambda\mu_* = \nabla L(\mu_*) = M\mathbf{d}$. In particular, $\nabla f(\mathbf{X}_{\mu_*}) = \text{diag}(\mathbf{d})\mathbf{X} = -\frac{2\lambda}{M}\mathbf{X}_{\mu_*}$. Thus the gradient is a scalar multiple of $\mathbf{X}_{\mu_*}$, and the tangent gradient to the Frobenius sphere vanishes. $\square$

The following lemma and Proposition III.2 show that Theorem V.1 applies to AWGN-ML-BLER, yielding Corollary V.6.

Lemma V.5.

Let $\mathbf{X}$ be a nondegenerate linear code with $M \geq 2$. No degenerate $\mathbf{X}_\mu$ locally minimizes AWGN-ML-BLER over $\mathcal{F}_{\mathbf{X}}$.

Proof Outline. Suppose μ deletes rows of $\mathbf{X}$, making $\mathbf{X}_\mu$ degenerate. Reallocate sufficiently small power to a non-constant deleted row. Remove duplicate codewords from $\mathbf{X}_\mu$ to obtain $\overline{\mathbf{X}}_\mu$ and use linearity of $\mathbf{X}$ to relate $P_c(\mathbf{X}_\mu)$ to $P_c(\overline{\mathbf{X}}_\mu)$. Differentiability of P_c at nondegenerate $\overline{\mathbf{X}}_\mu$ admits approximation of P_c for perturbations of μ , showing there is a small perturbation improving BLER. $\square$

Corollary V.6.

For AWGN-ML-BLER, for every nondegenerate linear code $\mathbf{X}$, it holds that $\mathbf{X}_{\mu_*}$ is a stationary point.

VI. STATIONARITY OF TRANSITIVE LINEAR CODES

We now show that for linear codes with a transitive automorphism group no reallocation is needed. The group $\text{Aut}(\mathbf{X})$ is *transitive* if for every coordinate pair $i, j \in [n]$ there exists $\mathbf{P} \in \text{Aut}(\mathbf{X})$ sending i to j .

Theorem VI.1.

Every nondegenerate linear code $\mathbf{X}$ with a transitive automorphism group is a stationary point.

Proof. Recall that $\mathbf{X}$ is stationary iff there exists $\lambda \in \mathbb{R}$ with $\nabla f(\mathbf{X}) + 2\lambda\mathbf{X} = 0$. To show this, we strengthen Corollary V.4 to $\text{Fix}(\text{Aut}_{\text{bi}}(\mathbf{X})) \subseteq \text{span}\{\mathbf{X}\}$. Since Lemma V.2 ensures $\nabla f(\mathbf{X}) \in \text{Fix}(\text{Aut}_{\text{bi}}(\mathbf{X}))$, we have stationarity.

Let $\mathbf{Y} \in \text{Fix}(\text{Aut}_{\text{bi}}(\mathbf{X}))$. By Corollary V.4, $\mathbf{Y} = \mathbf{X}_\mu$ for some $\mu \in \mathbb{R}^n$. By linearity, $\mathbf{x}_1 = 1$, so $\mu = \mathbf{Y}_{:,1}$. For $\mathbf{P} \in \text{Aut}(\mathbf{X})$ choose $\Pi \in S_M$ s.t. $(\mathbf{P}, \Pi) \in \text{Aut}_{\text{bi}}(\mathbf{X})$. Since $\mathbf{P}\mathbf{1} = 1$, reading off column 1 of $\mathbf{P}\mathbf{X}\Pi^{-1} = \mathbf{X}$ gives $\mathbf{P}\mathbf{x}_{\Pi^{-1}(1)} = \mathbf{x}_1 = 1$; thus we must have $\mathbf{x}_{\Pi^{-1}(1)} = 1$ as well. Note $\mathbf{Y} \in \text{Fix}(\text{Aut}_{\text{bi}}(\mathbf{X}))$, so $\mathbf{Y} = \mathbf{P}\mathbf{Y}\Pi^{-1}$; reading off column 1 on both sides:

$$\mu = \mathbf{P}\mathbf{Y}_{:, \Pi^{-1}(1)} = \mathbf{P} \text{diag}(\mu) \mathbf{x}_{\Pi^{-1}(1)} = \mathbf{P}\mu.$$

Since $\text{Aut}(\mathbf{X})$ acts transitively on $[n]$ and $\mathbf{P}\mu = \mu$ for every $\mathbf{P} \in \text{Aut}(\mathbf{X})$, $\mu = \alpha\mathbf{1}$ for some $\alpha \in \mathbb{R}$. Thus $\mathbf{Y} = \alpha\mathbf{X}$, and $\text{Fix}(\text{Aut}_{\text{bi}}(\mathbf{X})) \subseteq \text{span}\{\mathbf{X}\}$. $\square$

As both RM codes and cyclic codes are transitive [6], [30], we get the following corollaries.

Corollary VI.2.

- a) RM codes are stationary points.
- b) Cyclic codes are stationary points.

VII. ACKNOWLEDGMENT

Support from grants NSF 2217023 and NSF 2240532 is acknowledged. Support from Project 2024-1.2.3-HU-RIZONT-2024-00017 is acknowledged, financed by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, under the 2024-1.2.3-HU-RIZONT funding scheme. The authors acknowledge OpenAI for providing complimentary access to ChatGPT through the ChatGPT for Academic Researchers program. ChatGPT was used to support coding, literature review, and manuscript preparation.

REFERENCES

- [1] H. Kim, “Channel coding via machine learning,” in *Machine Learning and Wireless Communications*, Y. C. Eldar, A. Goldsmith, D. Gündüz, and H. V. Poor, Eds. Cambridge Univ. Press, 2022, pp. 77–109.
- [2] P. Petersen and J. Zech, “Mathematical theory of deep learning,” *CoRR*, vol. abs/2407.18384, 2024.
- [3] A. V. Makkuvu, X. Liu, M. V. Jamali, H. MahdaviFar, S. Oh, and P. Viswanath, “KO codes: inventing nonlinear encoding and decoding for reliable wireless communication via deep-learning,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 7368–7378.
- [4] Y. Jiang, H. Kim, H. Asnani, S. Kannan, S. Oh, and P. Viswanath, “Turbo autoencoder: Deep learning based channel codes for point-to-point communication channels,” in *NIPS*, Dec. 2019, pp. 2758–2768.
- [5] N. Devroye, A. Mulgund, R. Shekhar, G. Turán, M. Žefran, and Y. Zhou, “Interpreting training aspects of deep-learned error-correcting codes,” in *IEEE International Symposium on Information Theory, ISIT*. IEEE, 2023, pp. 2374–2379.
- [6] F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*. Amsterdam: North Holland, 1983.
- [7] S. Kudekar, S. Kumar, M. Mondelli, H. D. Pfister, E. Sasoglu, and R. L. Urbanke, “Reed–Muller Codes Achieve Capacity on Erasure Channels,” *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4298–4316, Jul. 2017.
- [8] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, 3rd ed. Springer, 1999.
- [9] H. Cohn and A. Kumar, “Universally optimal distribution of points on spheres,” *Journal of the American Mathematical Society*, vol. 20, no. 1, pp. 99–148, 2007.
- [10] G. Forney and G. Ungerboeck, “Modulation and coding for linear Gaussian channels,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2384–2415, 1998.
- [11] E. Tokarchuk, H. C. Bakker, and V. Niculae, “Keep your distance: learning dispersed embeddings on $\mathbb{S}_m$,” *Transactions on Machine Learning Research*, Apr. 2025.
- [12] W. C. Waterhouse, “Do symmetric problems have symmetric solution?” *The American Mathematical Monthly*, vol. 90, no. 6, pp. 378–387, 1983.
- [13] R. S. Palais, “The principle of symmetric criticality,” *Commun. Math. Phys.*, vol. 69, pp. 19–30, 1979.
- [14] M. Golubitsky and I. Stewart, *The Symmetry Perspective*. Basel: Birkhäuser Basel, 2002.
- [15] P. E. Hydon, “Discrete point symmetries of ordinary differential equations,” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1975, pp. 1961–1972, Jul. 1998.
- [16] —, *Symmetry Methods for Differential Equations: A Beginner’s Guide*, ser. Cambridge Texts in Applied Mathematics. Cambridge: Cambridge University Press, 2000, no. 22.
- [17] G. W. Bluman and S. C. Anco, *Symmetry and Integration Methods for Differential Equations*, ser. Applied Mathematical Sciences. New York, NY: Springer, 2002, vol. 154.
- [18] B. Simsek, F. Ged, A. Jacot, F. Spadaro, C. Hongler, W. Gerstner, and J. Brea, “Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances,” in *Proc. of the 38th International Conference on Machine Learning, ICML 2021*. PMLR, 2021, pp. 9722–9732.
- [19] Y. Xie and T. Smidt, “A Tale of Two Symmetries: Exploring the Loss Landscape of Equivariant Models,” in *Advances in Neural Information Processing Systems*, 2025, pp. 53 223–53 255.
- [20] D. Kunin, J. M. Bloom, A. Goeva, and C. Seed, “Loss landscapes of regularized linear autoencoders,” in *P. of the 36th International Conference on Machine Learning, ICML 2019*, ser. PMLR. PMLR, 2019, pp. 3560–3569.
- [21] N. Devroye, N. Mohammadi, A. Mulgund, H. Naik, R. Shekhar, G. Turán, Y. Wei, and M. Žefran, “Interpreting Deep-Learned Error-Correcting Codes,” in *ISIT*, Jun. 2022, pp. 2457–2462.
- [22] N. Devroye, A. Mulgund, R. Shekhar, G. Turán, Y. Wei, and M. Žefran, “Evaluating interpretations of deep-learned error-correcting codes,” in *Allerton*, Sep. 2022.
- [23] A. Mulgund, N. Devroye, G. Turán, and M. Žefran, “Decomposing the training of deep learned turbo codes via a feasible map decoder,” in *2023 12th International Symposium on Topics in Coding (ISTC)*, 2023, pp. 1–5.
- [24] Y. Zhou, N. Devroye, G. Turán, and M. Žefran, “Interpreting deepcode, a learned feedback code,” in *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 1403–1408.
- [25] Y. Zhou, N. Devroye, G. Turán, and M. Žefran, “Higher-order interpretations of deepcode, a learned feedback code,” in *2024 60th Annual Allerton Conference on Communication, Control, and Computing*, 2024, pp. 1–8.
- [26] Y. Zhou, N. Devroye, G. Turán, and M. Žefran, “On non-linearities of simple learned AWGN feedback codes,” in *2025 IEEE International Symposium on Information Theory (ISIT)*, 2025, pp. 1–6.
- [27] R. Shekhar, N. Devroye, G. Turán, and M. Žefran, “Interpreting KO codes,” in *2025 IEEE International Symposium on Information Theory (ISIT)*, 2025, pp. 1–6.
- [28] H. S. M. Coxeter, *Regular Polytopes*, 3rd ed. New York: Dover Publications, 1973.
- [29] N. Boumal, *An Introduction to Optimization on Smooth Manifolds*. Cambridge Univ. Press, 2023.
- [30] E. Abbe, O. Sberlo, A. Shpilka, and M. Ye, “Reed-muller codes,” *Found. Trends Commun. Inf. Theory*, vol. 20, no. 1-2, pp. 1–156, 2023.