Attribute-Efficient Evolvability of Linear Functions

Elaine Angelino
Harvard University
elaine@eecs.harvard.edu

Varun Kanade*
UC Berkeley
vkanade@eecs.berkeley.edu

November 3, 2013

Abstract

In a seminal paper, Valiant (2006) introduced a computational model for evolution to address the question of complexity that can arise through Darwinian mechanisms. Valiant views evolution as a restricted form of computational learning, where the goal is to *evolve* a hypothesis that is close to the *ideal function*. Feldman (2008) showed that (correlational) statistical query learning algorithms could be framed as evolutionary mechanisms in Valiant's model. P. Valiant (2012) considered evolvability of real-valued functions and also showed that weak-optimization algorithms that use weak-evaluation oracles could be converted to evolutionary mechanisms.

In this work, we focus on the *complexity* of representations of evolutionary mechanisms. In general, the reductions of Feldman and P. Valiant may result in intermediate representations that are arbitrarily complex (polynomial-sized circuits). We argue that biological constraints often dictate that the representations have low complexity, such as constant depth and fan-in circuits. We give mechanisms for evolving sparse linear functions under a large class of smooth distributions. These evolutionary algorithms are attribute-efficient in the sense that the size of the representations and the number of generations required depend only on the sparsity of the target function and the accuracy parameter, but have no dependence on the total number of attributes.

*This author is supported by a Simons Postdoctoral Fellowship.

1 Introduction

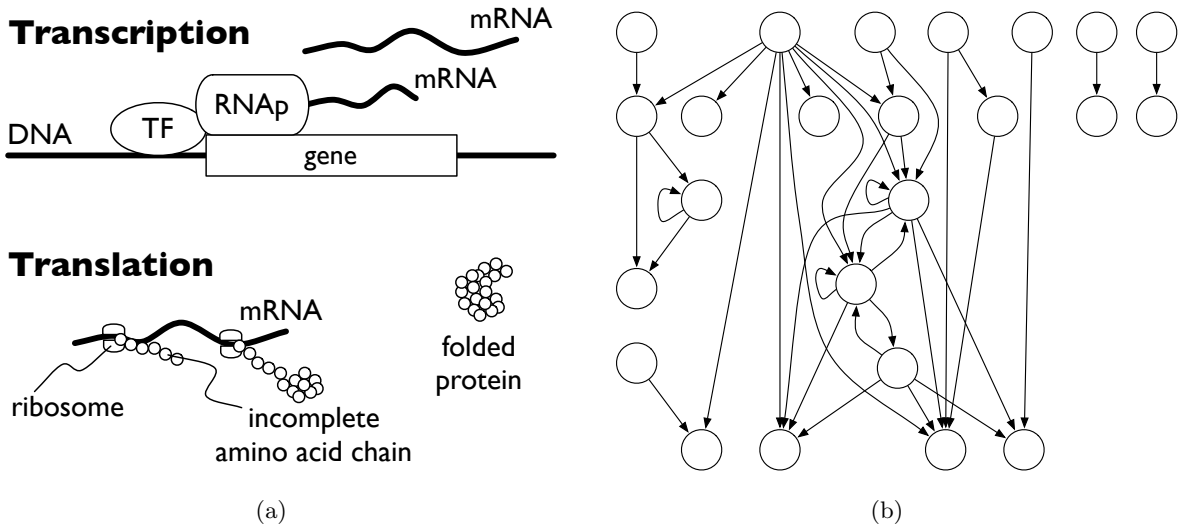
Darwin's theory of evolution through natural selection has been a cornerstone of biology for over a century and a half. Yet, a quantitative theory of complexity that could arise through Darwinian mechanisms has remained virtually unexplored. To address this question, Valiant introduced a computational model of evolution [28]. In his model, an organism is an entity that computes a function of its environment. There is a (possibly hypothetical) *ideal function* indicating the best behavior in every possible environment. The performance of the organism is measured by how close the function it computes is to the ideal. An organism produces a set of offspring, that may have mutations that alter the function computed. The performance (fitness) measure acting on a population of mutants forms the basis of natural selection. The resources allowed are the most generous while remaining feasible; the mutation mechanism may be any efficient randomized Turing machine, and the function represented by the organism may be arbitrary as long as it is computable by an efficient Turing machine.

Formulated this way, the question of evolvability can be asked in the language of computational learning theory. For what classes of ideal functions, C, can one expect to find an evolutionary mechanism that gets arbitrarily close to the ideal, within feasible computational resources? Darwinian selection is restrictive in the sense that the only feedback received is *aggregate* over life experiences. Valiant observed that any feasible evolutionary mechanism could be simulated in the statistical query framework of Kearns [19]. In a remarkable result, Feldman showed that in fact, evolvable concept classes are exactly captured by a restriction of Kearns' model, where the learning algorithm is only allowed to make *performance queries*, *i.e.*, it produces a hypothesis and then makes a query to an oracle that returns the (approximate) performance of that hypothesis under the distribution [9].[1] P. Valiant studied the evolvability of real-valued functions and showed that whenever the corresponding weak optimization problem, *i.e.*, approximately minimizing the expected loss, can be solved by using a weak evaluation oracle, such an algorithm can be converted into an evolutionary mechanism [29]. This implies that a large of class of functions – fixed-degree real polynomials – can be evolved with respect to any convex loss function.

Direct evolutionary mechanisms, not invoking the general reductions of Feldman and P. Valiant, have been proposed for certain classes in restricted settings. Valiant showed that the class of disjunctions is evolvable using a simple set of mutations under the uniform distribution [28]. Kanade, Valiant and Vaughan proposed a simple mechanism for evolving homogeneous linear separators under radially symmetric distributions [17]. Feldman considered a model where the ideal function is boolean but the representation can be real-valued, allowing for more detailed feedback. He presents an algorithm for evolving large margin linear separators for a large class of convex loss functions [11]. P. Valiant also showed that with very simple mutations, the class of fixed-degree polynomials can be evolved with respect to the squared loss [29].

Current understanding of biology (or lack thereof) makes it difficult to formalize a notion of *naturalness* for mutations in these frameworks; in particular, it is not well understood how mutations to DNA relate to functional changes in an organism. That said, the more direct algorithms are appealing due to the simplicity of their mutations. Also, the "chemical computers" of organisms may be slow, and hence, representations that have low complexity are attractive. In general, Feldman's generic reduction from statistical query algorithms may use arbitrarily complex representations (polynomial-sized circuits), depending on the specific algorithm used. In the remainder of the introduction, we first describe a particular class of biological circuits, *transcription networks*, that motivate our study. We then frame the evolutionary question in the language of computational

[1] Feldman calls these correlational statistical queries, because when working with boolean functions with range $\{-1, 1\}$, the performance of any hypothesis is its correlation with the ideal function.



(a) (b)

Figure 1: (a) Schematic of transcription (top) and translation (bottom). Here, a transcription factor (TF) binds to DNA close to a gene in a way that increases gene expression by encouraging RNA polymerase (RNAp) to transcribe the gene and so produce mRNA. The mRNA is then translated by ribosomes to produce sequences of amino acid that ultimately fold into proteins. Only a small number of transcription factors directly regulate any gene. Note that a transcription factor's action can also decrease gene expression. For a more complete picture, see *e.g.*, [1]. (b) Topology of the transcription network of respiration and redox reactions in yeast. $X \to Y$ represents that transcription factor X regulates the expression of Y. Note that this real network has cycles. Adapted from [23].

learning theory, summarize our contributions and discuss related work.

1.1 Representation in Biology

Biological systems appear to function successfully with greatly restricted representation classes. The nature of circuits found in biological systems may vary, but some aspects – such as *sparsity* – are common. Specifically, the interacting components in many biological circuits are sparsely connected. Biological circuits are often represented as networks or graphs, where the vertices correspond to entities such as neurons or molecules and the edges to connections or interactions between pairs of entities. For example, both neural networks [31] and networks of metabolic reactions in the cell [30, 14] have been described by "small-world" models, where a few "hub" nodes have many edges but most nodes have few edges (and consequently, the corresponding graphs have small diameter). An associated property observed in biological networks is *modularity*: a larger network of interacting entities is composed of smaller modules of (functionally related) entities [12]. Both the "small-world" description and modularity of biological networks are consistent with the more general theme of sparsity.

We focus on transcription networks, which are a specific class of networks of interacting genes and proteins that are involved in the production of new protein. Alon provides an accessible and mathematical introduction to transcription networks and other biological circuits [1]; below and in Figure 1(a), we present a simplified account that motivates this work. Genes are *transcribed* to produce mRNA, which is then *translated* into sequences of amino acids that ultimately fold

into proteins.[2] In a transcription network, a gene's transcription may be regulated by a set of proteins called *transcription factors*. These transcription factors may increase or decrease a gene's transcription by physically binding to regions of DNA that are typically close to the gene. In natural systems, only a small number of transcription factors regulate any single gene, and so transcription networks are sparsely connected. For example, Balaji *et al.* studied a yeast transcription network of 157 transcription factors regulating 4,410 genes. They observed this network to have 12,873 interactions (edges) where each gene was regulated on average by about 2.9 transcription factors, the distribution of in-degrees was well-described by an exponential fit, and only about 45 genes had an in-degree of 15 or greater [3].

The number of transcription factors varies from hundreds in a bacterium to thousands in a human cell. Some transcription factors are always present in the cell and can be thought of as representing a *snapshot* of the environment [1]. For example, the presence of sugar molecules in the environment may cause specific transcription factors to be *activated*, enabling them to regulate the production of other proteins. One of these proteins could be an *end-product*, such as an enzyme that catalyzes a metabolic reaction involving the sugar. Alternatively, the transcription factor could regulate another transcription factor that itself regulates other genes – we view this as intermediate computation – and may participate in further "computation" to produce the desired end-result.

While transcription networks may include cycles (loops), here for simplicity we focus on systems that are directed acyclic graphs, and the resulting computation can be viewed as a circuit. We illustrate a small, real transcription network in Figure 1(b). These circuits are by necessity shallow due to a temporal constraint, that the time required for sufficient quantities of protein to be produced is of the same order of magnitude as cell-division time.[3] For example, Luscombe *et al.* measured the shortest path length (in number of intermediate nodes) between transcription factors and regulated genes corresponding to terminal nodes (leaves) in a yeast transcription network. In the static network, the mean such path length was 4.7 and the longest path involved 12 intermediate transcription factors [21].

1.2 Our Contributions

First, our contribution is conceptual. We believe that the study of evolvability from a computational standpoint will benefit by understanding the representation complexity required to evolve a certain concept class. Motivated by the previous discussion, in the case of transcription networks, it appears essential that the representation used be a constant depth and fan-in (boolean or arithmetic) circuit. Of course, any function that can be represented by such a circuit can depend only on a constant number of input variables. We ask the question, when we restrict attention to functions in a given class that depend only on a constant number of variables, when can evolution succeed?

Second, we show that the class of sparse linear functions, those that depend only on a constant number of variables, under a large class of smooth distributions, can be evolved using sparse linear functions as representations, when the performance is measured using squared error. The number

[2] In reality, this is a dynamical system where the rates of production are important. Note that this process need not be linear: a gene (mRNA transcript) can be transcribed (translated) multiple times, not only in series but also in parallel fashion. We also ignore other *epigenetic* effects, *i.e.*, molecular modifications to DNA that do not change its sequence but alter gene expression, *e.g.*, the addition of methyl groups to nucleotides in a way that physically blocks transcription.

[3] Other kinds of networks, such as signaling networks, operate by changing the shapes of proteins. The fact that these transformations are rapid may allow for much larger depth. Note that fast conformational changes govern how transcription factors directly process information from the environment in order to regulate gene expression. In our example, a sugar molecule binds to a transcription factor and changes its shape in a way that alters its ability to bind to DNA.

of variables used by the representations is larger than the number of variables in the *ideal function* and depends on the *smoothness* parameter of the distribution. According to our notion of Δ-smooth G-nice distributions (Defn. 2), the density function of a smooth distribution is obtained by convolution of an arbitrary density with a product measure on $[-\sqrt{3}\Delta, \sqrt{3}\Delta]^n$ (alternatively, drawing a point from the smooth distribution is equivalent to drawing a point from an arbitrary distribution and adding a (noise) vector from a product distribution).

A linear function is represented by a weighted arithmetic circuit with only one addition gate (alternatively, by a depth-two circuit with a layer of multiplication gates and some constant inputs).[4] Also, the number of generations required for evolution to succeed depends polynomially on the sparsity k of the target linear function, the smoothness parameter Δ of the distribution and the inverse of the target accuracy ϵ, and has no dependence on the dimension n of the input space. Thus, our result shows *attribute-efficient* evolvability of sparse linear functions, in the sense of Littlestone [20]. For the precise statement, see Theorem 1 in Section 3.1.

Valiant also proposed a stronger selection mechanism – when natural selection aggressively selects the (almost) best mutation, rather than merely a beneficial one – called evolution by optimization. Our second result requires a much stronger distributional assumption – the correlation $\mathrm{corr}(x_i, x_j) \leq 1/(2k)$ – where k is the sparsity of the target linear function (see Defn. 3). Under such distributions, we show that under evolution by optimization, sparse linear functions can be evolved by representations with the same sparsity. The mechanism we propose and its analysis is inspired by the greedy orthogonal matching pursuit algorithms in signal processing [7, 27]. Unlike the previous evolutionary algorithm, this one requires initialization, *i.e.*, the evolutionary process begins with the 0 function. As in the previous case, the number of generations required depends polynomially on the sparsity k of the target linear function, the inverse of the accuracy parameter ϵ, but has no dependence on the total number of attributes n. The precise statement appears as Theorem 2 in Section 3.2.

Related Work

The question of proper vs. improper learning has been studied in computational learning theory. A separation between the two kinds is known, unless $\mathsf{NP} = \mathsf{RP}$. However, most interesting PAC-learnable classes can be learned using thresholds of low-degree polynomials, and do not seem to require the full generality of polynomial-sized circuits.[5] In this context, Valiant's disjunction algorithm under the uniform distribution [28], Kanade *et al.*'s algorithm for homogeneous half-spaces under radially symmetric distributions [17], and P. Valiant's algorithm for linear (polynomial) functions using squared loss [29], are *proper* evolutionary mechanisms, *i.e.*, the representation used is from the same class as the ideal function. In the first two cases, it is straightforward to show that if the target depends only on a constant number of variables, the evolutionary mechanism also succeeds using representations that depend only on a constant number of variables. Thus, attribute-efficient evolution can be achieved.

The problem of learning sparse linear functions has been studied under various names in several fields for many applications, *e.g.*, recovering sparse solutions to (underdetermined) linear systems of equations [4], or recovering sparse representations with a redundant dictionary [22, 8]; compressive sampling or compressed sensing for sparse signal reconstruction [5]; optimization with regularization or sparsity-inducing penalties in machine learning [2]; sparse coding for learning an overcomplete

[4] There is a natural tradeoff between fan-in and depth, that may be useful, depending on which is the more severe constraint.

[5] For example, the classes of k-CNF, k-term DNF, decision lists and low-rank decision trees, can all be represented as PTFs.

basis [25], or for denoising in image and video processing [8]. This area is too vast to review here; Bruckstein *et al.* have an excellent survey [4]. Learning the sparsest linear function is equivalent to finding the sparsest solution to a system of linear equations (assuming there is no noise in the data). In general, this problem is NP-hard and the currently best-known approximation factor depends on the norm of the pseudo-inverse of the matrix [24]. Thus, some assumption on the distribution seems necessary. Our evolution based on optimization algorithm (Section 3.2) is essentially the greedy orthogonal matching pursuit algorithm of Tropp [27] and Donoho *et al.* [7], cast in the language of evolvability; these algorithms are also known in statistical modeling as forward stepwise regression [6, 13].

Finally, the question of *attribute-efficient* regression in the PAC (or SQ) model is a natural one. Here, the goal would be to design a polynomial time algorithm for producing an ϵ-accurate linear function, with sample complexity that is polynomial in the sparsity k of the target function and the inverse of the target accuracy ϵ, and only polylogarithmic in n, the total number of attributes. Under mild boundedness assumptions on the distribution, this can be achieved by setting up an L_1-regularized optimization problem; the output classifier may not be sparse in light of the NP-hardness result mentioned above. We note that under the distributional assumption made in the paper, finding the *sparsest* linear function that fits the data is also easy in the PAC/SQ setting, since the solution to the optimization problem in this case is unique. The focus in our work is different, namely showing that simple evolutionary mechanisms can succeed, while using representations that are themselves sparse linear functions at all times.

Organization

In Section 2, we give an overview of Valiant's evolution model and describe the concept classes and class of distributions considered in this paper. Section 3 contains the mechanisms for evolving sparse linear functions. We conclude in Section 4 with some discussion and directions for future work.

2 Model and Preliminaries

We first provide an overview of the evolvability framework of Valiant [28]. The description here differs slightly from Valiant's original formulation and includes some subsequent extensions (for more details the reader is referred to [28, 9, 10, 29, 16]).

2.1 Valiant's Evolvability Framework

Let X denote a set of instances, *e.g.*, $X = \mathbb{R}^n$ or $X = \{0,1\}^n$. We assume that the representation length of each $x \in X$ is captured by the parameter n. To avoid excessive notation, we will keep this size parameter implicit in our description of the model. Let D be a distribution over X. Each $x \in X$ can be thought of as the description of an environmental setting, the inputs to any circuit of an organism. D denotes the distribution over the possible environmental settings an organism may experience in a lifetime. Let $f : X \to Y$ (typically $Y = \mathbb{R}$ or $Y = \{0,1\}$) denote the *ideal function*, the best behavior in each possible environmental setting.

Representations

A creature is a string representation that encodes an efficiently computable function $r : X \to Y$, *i.e.*, there is an efficient Turing Machine that, given the description string $\langle r \rangle$ and $x \in X$, outputs

$r(x)$.

In this work, our focus is characterizing different evolutionary mechanisms based on the complexity of representations used. The complexity of a representation is measured by the function it computes. Let $H : X \to Y$ be a class of functions. For $R \subseteq \{0,1\}^*$, we say that R *represents* H, if there is a map, $\sigma : R \to H$, and if there exists an *efficient* Turing machine that, given input $r \in R$ and $x \in X$, outputs $(\sigma(r))(x)$. Henceforth, by abuse of notation we will use r to denote both the representation and the function it computes, $\sigma(r)$.

Evolutionary Algorithms

The performance of a representation r is measured using a loss function $\ell : Y \times Y \to \mathbb{R}^+$, such that $\ell(y, y) = 0$. For a function $g : X \to Y$, define the expected loss with respect to the ideal function $f : X \to Y$, under distribution D, as $\mathrm{L}_{f,D}(g) = \mathbb{E}_{x \sim D}[\ell(g(x), f(x))]$.[6] The goal of evolution is to reach some representation r^* such that $\mathrm{L}_{f,D}(r^*) < \epsilon$. In the following discussion, we use the notation: f the ideal function, ϵ the target accuracy, D the target distribution over X and $\mathrm{L}_{f,D}(g)$ the expected loss function.

Mutator: A mutator $\mathsf{Mut}(r, \epsilon)$, for a set of representations R, is a polynomial-time randomized Turing machine that takes as input a representation $r \in R$ and accuracy parameter ϵ and outputs a multiset $\mathsf{Neigh}(r, \epsilon) \subseteq R$. The running time requirement on Mut also ensures that $|\mathsf{Neigh}(r, \epsilon)|$ is polynomially bounded.

Selection: (Natural) Selection is based on the empirical performance of each representation. Let $s : R \times [0,1] \to \mathbb{N}$ be a sample size function. First, the mutation algorithm, $\mathsf{Mut}(r, \epsilon)$, is run to produce multiset $\mathsf{Neigh}(r, \epsilon)$. Then, an i.i.d. sample $\langle x^i \rangle_{i=1}^{s}$ is drawn from the distribution D over X, where $s = s(r, \epsilon)$. Denote the empirical performance of each $r' \in \mathsf{Neigh}(r, \epsilon) \cup \{r\}$ as

$$\hat{\mathrm{L}}_{f,D}(r') = \frac{1}{s} \sum_{i=1}^{s} \ell(r'(x^i), f(x^i))$$

Finally, let $t : R \times [0,1] \to \mathbb{R}$ be a tolerance function. Two possible selection mechanisms are considered.

1. **Selection based on beneficial and neutral mutations** (BN-Sel): Let

   $$\mathsf{Bene} = \{r' \in \mathsf{Neigh}(r, \epsilon) \mid \hat{\mathrm{L}}_{f,D}(r') \leq \hat{\mathrm{L}}_{f,D}(r) - t(r, \epsilon)\}$$

   denote the set of beneficial mutations and let

   $$\mathsf{Neut} = \{r' \in \mathsf{Neigh}(r, \epsilon) \mid |\hat{\mathrm{L}}_{f,D}(r') - \hat{\mathrm{L}}_{f,D}(r)| < t(r, \epsilon)\}$$

   denote the neutral mutations, with respect to tolerance function t. Both Bene and Neut are treated as multisets (the multiplicity of any representation is the same as that in $\mathsf{Neigh}(r, \epsilon)$). Selection operates as follows: if $\mathsf{Bene} \neq \emptyset$, r' is randomly selected from Bene as the surviving creature at the next generation. If $\mathsf{Bene} = \emptyset$ and $\mathsf{Neut} \neq \emptyset$, then r' is selected randomly from Neut as the surviving creature at the next generation. Otherwise, $\perp$ is produced signifying failure of evolution.

[6] This definition does not require the expected loss to be bounded, but we will mainly be interested in situations when that is the case.

2. **Selection based on optimization** (Opt-Sel): Let $\widehat{\mathsf{opt}} = \min_{r' \in \mathsf{Neigh}(r,\epsilon)} \hat{\mathrm{L}}_{f,D}(r')$. If $\widehat{\mathsf{opt}} > \hat{\mathrm{L}}_{f,D}(r) + t(r,\epsilon)$, then $\perp$ is produced signifying failure of evolution. Otherwise, consider the multiset, $\mathsf{Best} = \{r' \in \mathsf{Neigh}(r,\epsilon) \mid \hat{\mathrm{L}}_{f,D}(r') \leq \widehat{\mathsf{opt}} + t(r,\epsilon)\}$, and then r' is chosen from Best randomly as the surviving creature at the next generation.

Thus, while the selection rule BN-Sel only chooses some beneficial (or at least neutral) mutation, Opt-Sel aggressively picks the (almost) best mutation from the available pool.

We denote by $r' \leftarrow \mathsf{Sel}[R, \mathsf{Mut}, s, t](r, \epsilon)$ the fact that r' is the surviving creature in the next generation after one mutation and selection operation on the representation r and accuracy parameter ϵ. Here, Sel may be one of the two selection rules described above. For Sel to be feasible we require that the size function s is polynomially bounded (in n and $1/\epsilon$) and that the inverse of the tolerance function t is polynomially sandwiched, *i.e.*, there exists polynomials $p_1(n, 1/\epsilon)$ and $p_2(n, 1/\epsilon)$ such that $1/p_1(n, 1/\epsilon) \leq t(r, \epsilon) \leq 1/p_2(n, 1/\epsilon)$ for every $r \in R$ and $\epsilon > 0$.

Evolutionary Algorithm: An evolutionary algorithm $\mathcal{E}A$ is a tuple $(R, \mathsf{Mut}, s, t, \mathsf{Sel})$. When $\mathcal{E}A$ is run starting from $r_0 \in R$ with respect to distribution D over X, ideal function $f : X \to Y$, loss function ℓ and parameter ϵ, a sequence $r_0, r_1, r_2, \ldots$ is produced, where $r_i \leftarrow \mathsf{Sel}[R, \mathsf{Mut}, s, t](r_{i-1}, \epsilon)$. If $r_i = \perp$ for some i, we consider evolution as halted and $r_j = \perp$ for $j > i$. We say that $\mathcal{E}A$ succeeds at generation g, if g is the smallest index for which the expected loss $\mathrm{L}_{f,D}(r_g) \leq \epsilon$.

Definition 1 (Evolvability [28])**.** *We say that a concept class C is evolvable with respect to loss function ℓ and selection rule* Sel*, under a class of distributions $\mathcal{D}$ using a representation class H, if there exists a representation scheme $R \subseteq \{0,1\}^*$, such that R represents H, and there exists an evolutionary algorithm $\mathcal{E}A = (R, \mathsf{Mut}, s, t, \mathsf{Sel})$, such that for every $D \in \mathcal{D}$, every $f \in C$, every $\epsilon > 0$, and every $r_0 \in R$, with probability at least $1 - \epsilon$, $\mathcal{E}A$ run starting from r_0 with respect to f, D, ℓ, ϵ, produces r_g for which $\mathrm{L}_{f,D}(r_g) < \epsilon$. Furthermore, the number of generations g required for evolution to succeed should be bounded by a polynomial in n and $1/\epsilon$.*

Remark 1. *If the evolutionary algorithm succeeds only for a specific starting representation r_0, we say C is evolvable with* initialization*.*

Remark 2. *If the functions in concept class C depend only on k variables, we say the evolutionary algorithm is attribute-efficient, if the size function, s, is polylogarithmic in n and polynomial in k and $1/\epsilon$ and the number of generations, g, is polynomial in k and $1/\epsilon$ and does not depend on n.*

The definition presented above varies slightly from the definition of Valiant, in the sense that we explicitly focus on the complexity of representations used by the evolutionary algorithm. As discussed in the introduction, we focus on concept classes where each function depends on *few* (constant) input variables.[7]

2.2 Sparse Linear Functions

Our main result in this paper concerns the class of sparse linear functions. We represent a linear function from $\mathbb{R}^n \to \mathbb{R}$ by a vector $w \in \mathbb{R}^n$, where $x \mapsto w \cdot x$. For a vector $w \in \mathbb{R}^n$, sparsity(w) is the number of non-zero elements of w.

For any $0 \leq l < u$ and integer k, define the class of linear functions:

$$\mathsf{Lin}^k_{l,u} = \{x \mapsto w \cdot x \mid \text{sparsity}(w) \leq k, \forall i, w_i = 0 \text{ or } l \leq |w_i| \leq u\}$$

[7]These functions have been referred to as juntas in the theory literature. We avoid using this nomenclature as we restrict our attention to specific functional forms, such as linear functions, with k relevant variables.

Thus, $\mathsf{Lin}^k_{l,u}$ is the class of k-sparse linear functions, where the "influence" of each variable is upper and lower bounded.[8]

Let D be a distribution over $\mathbb{R}^n$. For $w, w' \in \mathbb{R}^n$, define the inner product $\langle w, w' \rangle = \mathbb{E}_{x \sim D}[(w \cdot x)(w' \cdot x)]$, where $w \cdot x = \sum_{i=1}^n w_i x_i$ denotes the standard dot product in $\mathbb{R}^n$. In this paper, we use $\|w\|$ to denote $\sqrt{\langle w, w \rangle}$ (and not $\sqrt{\sum_i w_i^2}$). To avoid confusion, whenever necessary, we will refer to the quantity $\sqrt{\sum_i w_i^2}$ explicitly if we mean the standard Euclidean norm.

Distribution Classes

We use two classes of distributions for our results in this paper. We define them formally here.

Smooth Bounded Distributions: We consider the class of smooth bounded distributions over $\mathbb{R}^n$. The concept of smoothed analysis of algorithms was introduced by Spielman and Teng [26] and recently the idea has been used in learning theory [15, 18]. We consider distributions that are bounded and have 0 mean. Formally, distributions we consider are defined as:

Definition 2 (Δ-Smooth G-Nice Distribution)**.** *A distribution D is a Δ-smooth G-nice distribution if it is obtained as follows. Let $\tilde{D}$ be some distribution over $\mathbb{R}^n$, and let U_a^n denote the uniform distribution over $[-a, a]^n$. Then $D = \tilde{D} * U^n_{\sqrt{3}\Delta}$ is obtained by the convolution of $\tilde{D}$ with $U^n_{\sqrt{3}\Delta}$.*[9] *Furthermore, D satisfies the following:*

1. $\mathbb{E}_D[x] = 0$
2. *For all i,* $\mathbb{E}_D[x_i^2] \leq 1$
3. *For every x in the support of D,* $\sum_{i=1}^n x_i^2 \leq G^2$

Incoherent Distributions: We also consider *incoherent* distributions.[10] For a distribution D over $\mathbb{R}^n$, the coherence is defined as $\max_{i,j} \operatorname{corr}(x_i, x_j)$, where $\operatorname{corr}(x_i, x_j)$ is the correlation between x_i and x_j. Again, we consider bounded distributions with zero mean. We also require the variance to be upper and lower bounded in each dimension. Formally, the distributions we consider are defined as:

Definition 3 (μ-Incoherent (Δ, G)-Nice Distribution)**.** *A distribution D is a μ-incoherent (Δ, G)-nice distribution if the following hold:*

1. $\mathbb{E}_D[x] = 0$
2. *For all i,* $\Delta^2 \leq \mathbb{E}_D[x_i^2] \leq 1$
3. *For all i, j,* $\max_{i,j} \operatorname{corr}(x_i, x_j) \leq \mu$
4. *For all x in the support of D,* $\sum_{i=1}^n x_i^2 \leq G^2$

[8] We do not use the word "influence" in the precise technical sense here.

[9] We could perform convolution with a spherical Gaussian distribution, however, this would make the resulting distribution unbounded. All results in this paper hold if we work with sub-Gaussian distributions and consider convolution with a spherical Gaussian distribution with variance Δ^2. In this case, we would be required to use Chebychev's inequality rather than Hoeffding's bound to show that the empirical estimate is close to the expected loss with high probability.

[10] This terminology is adapted from incoherence of matrices, *e.g.*, see [4].

We say a linear function represented by $w \in \mathbb{R}^n$ is W-bounded if $\sum_{i=1}^n w_i^2 \leq W^2$. We use the notation $w(x) = w \cdot x$. Suppose f, w are W-bounded linear functions, and distribution D is such that for every x in the support of D, $\sum_{i=1}^n x_i^2 \leq G^2$. We consider the squared loss function, which for $y, y' \in \mathbb{R}$ is $\ell(y', y) = (y' - y)^2$. Then, for any x in the support of D, $\ell(f(x), w(x)) \leq 4W^2G^2$. Thus, standard Hoeffding bounds imply that if $\langle x^i \rangle_{i=1}^s$ is an i.i.d. sample drawn from D, then

$$\Pr\left[\left|\frac{1}{s}\sum_{i=1}^{s} \ell(w(x^i), f(x^i)) - \mathrm{L}_{f,D}(w)\right| \geq \tau\right] \leq 2\exp\left(-\frac{s\tau^2}{8W^2G^2}\right) \tag{1}$$

Finally, for linear functions w ($x \mapsto w \cdot x$), let $\mathsf{NZ}(w) = \{i \mid w_i \neq 0\}$ denote the non-zero variables in w, so sparsity$(w) = |\mathsf{NZ}(w)|$. Then, we have the following Lemma. The proof appears in Appendix A.1.

Lemma 1. *Let D be a Δ-smooth G-nice distribution (Defn. 2), let $w \in \mathbb{R}^n$ be a vector and consider the corresponding linear function, $x \mapsto w \cdot x$. Then the following are true:*

1. *For any $1 \leq i \leq n$, $w_i^2 \leq \frac{\langle w,w \rangle}{\Delta^2}$.*
2. *There exists an i such that $w_i^2 \leq \frac{\langle w,w \rangle}{|\mathsf{NZ}(w)|\Delta^2}$.*

3 Evolving Sparse Linear Functions

In this section, we describe two evolutionary algorithms for evolving sparse linear functions. The first evolves the class $\mathsf{Lin}_{l,u}^k$ under the class of Δ-smooth G-nice distributions (Defn. 2), using the selection rule BN-Sel. The second evolves the class $\mathsf{Lin}_{0,u}^k$ under the more restricted class of $(1/2k)$-incoherent (Δ, G)-nice distributions (Defn. 3), using the selection rule Opt-Sel. We first define the notation used in the rest of this section.

Notation: D denotes the target distribution over $X = \mathbb{R}^n$, f denotes the ideal (target) function. The inner product $\langle \cdot, \cdot \rangle$ and 2-norm $\| \cdot \|$ of functions are with respect to the distribution D. $[n]$ denotes the set $\{1, \ldots, n\}$. For $S \subseteq [n]$, f^S denotes the best linear approximation of f using the variables in the set S; formally,

$$f^S = \operatorname*{argmin}_{w \in \mathbb{R}^n \,:\, w_i = 0 \,\vee\, i \in S} \|f - w\|^2$$

Finally, recall that for $w \in \mathbb{R}^n$, $\mathsf{NZ}(w) = \{i \mid w_i \neq 0\}$ and sparsity$(w) = |\mathsf{NZ}(w)|$. A vector w represents a linear function, $x \mapsto w \cdot x$. The vector e^i has 1 in coordinate i and 0 elsewhere and corresponds to the linear function $x \mapsto x_i$. Thus, in this notation, $\mathrm{corr}(x_i, x_j) = \langle e^i, e^j \rangle / (\|e^i\| \|e^j\|)$. The accuracy parameter is denoted by ϵ.

3.1 Evolving Sparse Linear Functions Using BN-Sel

We present a simple mechanism that evolves the class of sparse linear functions $\mathsf{Lin}_{l,u}^k$ with respect to Δ-smooth G-nice distributions (see Defn. 2). The representation class also consists of sparse linear functions, but with a greater number of non-zero entries than the *ideal function*. We also assume that a linear function is represented by $w \in \mathbb{R}^n$, where each w_i is a real number. (Handling the issues of finite precision is standard and is avoided in favor of simplicity.) Define the parameters $K = 5184(k/\Delta)^4(u/l)^2$ and $B = 10uk/\Delta$. Formally, the representation class is:

$$R = \{w \mid \text{sparsity}(w) \leq K, w_i \in [-B, B]\}$$

The important point to note is that the parameters K and B do not depend on n, the total number of variables.

Next, we define the mutator. Recall that the mutator is a randomized algorithm that takes as input an element $r \in R$ and accuracy parameter ϵ, and outputs a multiset $\mathsf{Neigh}(r, \epsilon) \subseteq R$. Here, $\mathsf{Neigh}(r, \epsilon)$ is populated by m independent draws from the following procedure, where m will be specified later (see the proof of Theorem 1). Starting from $w \in R$, define the mutated representation w', output by the mutator, as:

1. *Scaling*: With probability 1/3, choose $\gamma \in [-1, 1]$ uniformly at random and let $w' = \gamma w$.
2. *Adjusting*: With probability 1/3, do the following. Pick $i \in \mathsf{NZ}(w) = \{i \mid w_i \neq 0\}$ uniformly at random. Let w' denote the mutated representation, where $w'_j = w_j$ for $j \neq i$, and choose $w'_i \in [-B, B]$ uniformly at random.
3. With the remaining 1/3 probability, do the following:
   (a) *Swapping*: If $|\mathsf{NZ}(w)| = K$, choose $i_1 \in \mathsf{NZ}(w)$ uniformly at random. Then, choose $i_2 \in [n] \setminus \mathsf{NZ}(w)$ uniformly at random. Let w' be the mutated representation, where $w'_j = w_j$ for $j \neq i_1, i_2$. Set $w'_{i_1} = 0$ and choose $w'_{i_2} \in [-B, B]$ uniformly at random. In this case, $\text{sparsity}(w') = \text{sparsity}(w) = K$ with probability 1, and hence $w' \in R$.
   (b) *Adding*: If $|\mathsf{NZ}(w)| < K$, choose $i \in [n] \setminus \mathsf{NZ}(w)$ uniformly at random. Let w' be the mutated representation, where $w'_j = w_j$ for $j \neq i$, and choose $w'_i \in [-B, B]$ uniformly at random.

Recall that $f \in \mathsf{Lin}^k_{l,u}$ denotes the ideal (target) function and D is the underlying distribution that is Δ-smooth G-nice (see Defn. 2). Since we are working with the squared loss metric, $\ell(y', y) = (y' - y)^2$, the expected loss for any $w \in R$ is given by $\mathrm{L}_{f,D}(w) = \|f - w\|^2 = \langle f - w, f - w \rangle$. We will show that for any $w \in R$, if $\|f - w\|^2 > \epsilon$, with non-negligible (inverse polynomial) probability, the above procedure produces a mutation w' that decreases the expected loss by at least some inverse polynomial amount. Thus, by setting the size of the neighborhood m large enough, we can guarantee that with high probability there will always exist a beneficial mutation.

To simplify notation, let $S = \mathsf{NZ}(w)$. Recall that f^S denotes the best approximation to f using variables in the set S; thus, $\|f - w\|^2 = \|f - f^S\|^2 + \|f^S - w\|^2$. At a high level, the argument for proving the success of our evolutionary mechanism is as follows: If $\|f^S - w\|^2$ is large, then a mutation of the type "scaling" or "adjusting" will get w closer to f^S, reducing the expected loss. (The role of "scaling" mutations is primarily to ensure that the representations remain bounded.) If $\|f^S - w\|^2$ is small and $S \neq \mathsf{NZ}(f)$, there must be a variable in $\mathsf{NZ}(f) \setminus S$, that when added to w (possibly by swapping), reduces the expected loss. Thus, as long as the representation is far from the evolutionary target, a *beneficial* mutation is produced with high probability.

More formally, let w' denote a random mutation produced as a result of the procedure described above. We will establish the desired result by proving the following claims.

Claim 1. *If $\|w\| \geq 2\|f^S\|$, then with probability at least 1/12, $\mathrm{L}_{f,D}(w') \leq \mathrm{L}_{f,D}(w) - \|f^S - w\|^2/12$. In particular, a "scaling" type mutation achieves this.*

Claim 2. *When $\|w\| \leq 2\|f^S\|$, then with probability at least $\Delta\|f^S - w\|/(6K^2B)$, $\mathrm{L}_{f,D}(w') \leq \mathrm{L}_{f,D}(w) - 3\Delta^2\|f^S - w\|^2/(4|S|^2)$. In particular, an "adjusting" type mutation achieves this.*

Claim 3. *When $\|f^S - w\| \leq l^2\Delta^2/(4KB)$, but $\mathsf{NZ}(f) \not\subseteq S$, then with probability at least $\Delta\|f - w\|/(6KBnk)$, $\mathrm{L}_{f,D}(w') \leq \mathrm{L}_{f,D}(w) - \Delta^2\|f - w\|^2/(16k^2)$. In particular, a mutation of type "swapping" or "adding" achieves this.*

Note that when $\mathsf{NZ}(f) \subseteq S$, then $f^S = f$. Thus, in this case when $\mathrm{L}_{f,D}(w) = \|f^S - w\|^2 \leq \epsilon$, the evolutionary algorithm has succeeded.

The proofs of the above Claims are provided in Appendix A.2. We now prove our main result using the above claims.

Theorem 1. *Let $\mathcal{D}$ be the class of Δ-smooth G-nice distributions over $\mathbb{R}^n$ (Defn. 2). Then the class $\mathsf{Lin}^k_{l,u}$ is evolvable with respect to $\mathcal{D}$, using the representation class $\mathsf{Lin}^K_{0,B}$, where $K = O((k/\Delta)^4(u/l)^2)$ and $B = O(uk/\Delta)$, using the mutation algorithm described in this section, and the selection rule* BN-Sel. *Furthermore, the following are true:*

1. *The number of generations required is polynomial in (u/l), $1/\epsilon$, $1/\Delta$, and is independent of n, the total number of attributes.*
2. *The size function s, the number of points used to calculate empirical losses, depends polylogarithmically on n, and polynomially on the remaining parameters.*

Proof. The mutator is as described in this section. Let

$$p = \min\left\{\frac{1}{12}, \frac{l^2\Delta^3}{24K^3B^2}, \frac{\Delta\sqrt{\epsilon}}{6KBnk}\right\},$$

and let

$$\alpha = \min\left\{\frac{l^4\Delta^4}{192K^2B^2}, \frac{3l^4\Delta^6}{64K^4B^2}, \frac{\epsilon\Delta^2}{16k^2}\right\}.$$

Now, by Claims 1, 2 and 3, if $\|f - w\|^2 \geq \epsilon$, then the mutator outputs a mutation that decreases the squared loss by α with probability at least p.

Recall that $K = 5184(k/\Delta)^4(u/l)^2$ and $B = 10uk/\Delta$. Now, let $g = 20KG^2B^2/\alpha$ (recall that G^2 is the bound on $\sum_i x_i^2$ for x in the support of the distribution). We will show that evolution succeeds in at most g generations. Note that g has no dependence on n, the number of attributes, and polynomial dependence on the remaining parameters. Define $m = p^{-1}\ln(2g/\epsilon)$, and at each time step we have that $|\mathsf{Neigh}(w, \epsilon)| = m$. Note that together with the observation above, this implies that except with probability $\epsilon/2$, for $1 \leq i \leq g$, if w^i is the representation at time step i, $\mathsf{Neigh}(w^i, \epsilon)$ contains a mutation that decreases the loss by at least α, if $\mathrm{L}_{f,D}(w^i) \geq \epsilon$.

Now, let $t = 3\alpha/5$ be the *tolerance function*, set $\tau = \alpha/5$ and let $s = (200KG^2B^2/\alpha^2)\ln(4mg/\epsilon)$ be the *size function*. Note that $\sum_i w_i^2 \leq KB^2$ for $w \in R$ (this also holds for f, since $k < K$ and $u < B$). If $\langle x^i\rangle_{i=1}^s$ is an i.i.d. sample drawn from D, for each $\bar{w}$ of the mg representations that may be considered in the neighborhoods for the first g time steps, using (1), it holds that $|\hat{\mathrm{L}}_{f,D}(\bar{w}) - \mathrm{L}_{f,D}(\bar{w})| \leq \tau$ simultaneously except with probability $\epsilon/2$ (by a union bound). Thus, allowing for failure probability ϵ, we assume that we are in the case when the neighborhood always has a mutation that decreases the expected loss by α (whenever the expected loss of the current representation is at least ϵ) and that all empirical expected losses are τ-close to the true expected losses.

Now let w be the representation at some generation such that $\mathrm{L}_{f,D}(w) \geq \epsilon$, let $w' \in \mathsf{Neigh}(w, \epsilon)$ such that $\mathrm{L}_{f,D}(w') \leq \mathrm{L}_{f,D}(w) - \alpha$. Then, it is the case that $\hat{\mathrm{L}}_{f,D}(w') \leq \hat{\mathrm{L}}_{f,D}(w) - 3\alpha/5$ (when $\tau = \alpha/5$). Hence, for tolerance function $t = 3\alpha/5$, for the selection rule using BN-Sel, $w' \in \mathsf{Bene}$. Consequently $\mathsf{Bene} \neq \emptyset$. Hence, the representation at the next generation is chosen from Bene. Let $\tilde{w}$ be the chosen representation. It must be the case that $\hat{\mathrm{L}}_{f,D}(\tilde{w}) \leq \hat{\mathrm{L}}_{f,D}(w) - t$. Thus, we have $\mathrm{L}_{f,D}(\tilde{w}) \leq \mathrm{L}_{f,D}(w) - \alpha/5$. Hence, the expected loss decreases at least by $\alpha/5$.

Note that at no point can the expected loss be greater than $4KG^2B^2$ for any representation in R. Hence, in at most $20KG^2B^2/\alpha$ generations, evolution reaches a representation with expected loss at most ϵ. Note the only parameter introduced which has an inverse polynomial dependence on n is p. This implies that s only has polylogarithmic dependence on n. This concludes the proof of the theorem. □

Remark 3. *We note that the same evolutionary mechanism works when evolving the class* $\mathsf{Lin}^k_{0,u}$*, as long as the sparsity* K *of the representation class is allowed polynomial dependence on* $1/\epsilon$*, the inverse of the accuracy parameter. This is consistent with the notion of attribute-efficiency, where the goal is that the information complexity should be polylogarithmic in the number of attributes, but may depend polynomially on* $1/\epsilon$*.*

3.2 Evolving Sparse Linear Functions Using Opt-Sel

In this section, we present a different evolutionary mechanism for evolving sparse linear functions. This algorithm essentially is an adaptation of a greedy algorithm commonly known as orthogonal matching pursuit (OMP) in the signal processing literature (see [7, 27]). Our analysis requires stronger properties on the distribution: we show that k-sparse linear functions can be evolved with respect to $1/(2k)$-incoherent (Δ, G)-nice distributions (Defn. 3). Here, the selection rule used is selection using *optimization* (Opt-Sel).[11] Also, the algorithm is guaranteed to succeed only with *initialization* from the 0 function. Nevertheless, this evolutionary algorithm is appealing due to its simplicity and because it never uses a representation that is not a k-sparse linear function.

Recall that $f \in \mathsf{Lin}^k_{0,u}$ is the ideal (target) function.[12] Let

$$R = \{w \mid \text{sparsity}(w) \leq k, w_i \in [-B, B]\},$$

where $B = 10uk/\Delta$. Now, starting from $w \in R$, define the action of the mutator as follows (we will define the parameters λ and m later in the proof of Theorem 2):

1. *Adding*: With probability λ, do the following. Recall that $\mathsf{NZ}(w)$ denotes the non-zero entries of w. If $|\mathsf{NZ}(w)| < k$, choose $i \in [n] \setminus \mathsf{NZ}(w)$ uniformly at random. Let w' be such that $w'_j = w_i$ for $j \neq i$, and choose $w'_i \in [-B, B]$ uniformly at random. If $\mathsf{NZ}(w) = k$, let $w' = w$. Then, the multiset $\mathsf{Neigh}(w, \epsilon)$ is populated by m independent draws from the procedure just described.

2. With probability $1 - \lambda$, do the following:

   (a) *Identical*: With probability $1/2$, output $w' = w$.

   (b) *Scaling*: With probability $1/4$, choose $\gamma \in [-1, 1]$ uniformly at random and let $w' = \gamma w$.

   (c) *Adjusting*: With probability $1/4$, do the following. Pick $i \in \mathsf{NZ}(w)$ uniformly at random. Let w' be such that $w'_j = w_j$ for $j \neq i$, and choose $w'_i \in [-B, B]$ uniformly at random.

   Then, the multiset $\mathsf{Neigh}(w, \epsilon)$ is populated by m independent draws from the procedure just described.

[11] Valiant showed that selection using optimization was equivalent to selection using beneficial and neutral mutations [28]. However, this reduction uses representation classes that may be somewhat complex. For restricted representation classes, it is not clear that such a reduction holds. In particular, the necessary ingredient seems to be *polynomial-size* memory.

[12] Here, we no longer need the fact that each coefficient in the target linear function has magnitude at least l.

One thing to note in the above definition is that the mutations produced by the mutator at any given time are correlated, *i.e.*, they are all either of the kind that add a new variable, or all of the kind that just manipulate existing variables. At a high level, we prove the success of this mechanism as follows:

1. Using mutations of type "scaling" or "adjusting," a representation that is close to the *best* in the space, *i.e.*, f^S, is evolved.

2. When the representation is (close to) the best possible using current variables, adding one of the variables that is present in the *ideal function*, but not in the current representation, results in the greatest reduction of expected loss. Thus, selection based on optimization would always add a variable in $\mathsf{NZ}(f)$. By tuning λ appropriately, it is ensured that with high probability, candidate mutations that add new variables are not chosen until evolution has had time to approach the *best* representation using existing variables.

To complete the proof we establish the following claims.

Claim 4. *If $\|f^S - w\| \leq \sqrt{\epsilon}/2k$, then if $S \subsetneq \mathsf{NZ}(f)$, there exists $i \in \mathsf{NZ}(f) \setminus S$ and $-B < a < b < B$, such that for any $\gamma \in [a, b]$, $\mathrm{L}_{f,D}(w+\gamma e^i) \leq \mathrm{L}_{f,D}(w) - \epsilon/(4k^2)$ and for any $j \notin \mathsf{NZ}(f)$, $\beta \in [-B, B]$, $\mathrm{L}_{f,D}(w + \beta e^j) \geq \mathrm{L}_{f,D}(w + \gamma e^i) + \epsilon/(4k^3)$. Furthermore, $b - a \geq \sqrt{(k+1)\epsilon}/k^2$.*

Claim 5. *Conditioned on the mutator not outputting mutations that add a new variable, with probability at least $\min\{1/16, \|f^S - w\|/(16k^2B)\}$, there exists a mutation that reduces the squared loss by at least $\|f^S - w\|^2/(12k^2)$.*

The proofs of Claims 4 and 5 are not difficult and are provided in Appendix A.3. Based on the above claims we can prove the following theorem:

Theorem 2. *Let $\mathcal{D}$ be the class of $1/(2k)$-incoherent (Δ, G)-nice distributions over $\mathbb{R}^n$ (Defn. 3). Then, the class $\mathsf{Lin}^k_{0,u}$ is evolvable with respect to $\mathcal{D}$ by an evolutionary algorithm, using the mutation algorithm described in this section, selection rule $\mathsf{Opt\text{-}Sel}$, and the representation class $R = \mathsf{Lin}^k_{0,B}$, where $B = 10uk/\Delta$. Furthermore, the following are true:*

1. *The number of generations g is polynomial in $1/\epsilon$, k, $1/\Delta$, but independent of the dimension n.*

2. *The size function s, the number of points used to calculate the empirical losses, depends polylogarithmically on n and polynomially on the remaining parameters.*

Proof. The proof is straightforward, although a bit heavy on notation; we provide a sketch here. The mutator is as described in this section. Let

$$p = \min\left\{\frac{1}{16}, \frac{\sqrt{\epsilon}}{64k^3B}, \frac{\sqrt{(k+1)\epsilon}}{k^2}\right\},$$

and let

$$\alpha = \min\left\{\frac{\epsilon}{4k^3}, \frac{\epsilon}{192k^4}\right\} = \frac{\epsilon}{192k^4}.$$

Also, let $\tau = \alpha/5$ and let $t = 3\alpha/5$ be the *tolerance function*.

First, we show that between the "rare" time steps when the mutator outputs mutations that add a new variable, evolution has enough time to *stabilize* (reach close to local optimal) using existing variables. To see this, consider a sequence of coin tosses, where the probability of heads is λ and the probability of tails is $1 - \lambda$. Let Y_i be the number of tails between the $(i-1)^{th}$ and i^{th}

heads. Except with probability $\epsilon/(4(k+1))$, $Y_i > \epsilon/(4(k+1)\lambda)$ by a simple union bound. Also, by Markov's inequality, except with probability $\epsilon/(4(k+1))$, $Y_i < 4(k+1)/(\epsilon\lambda)$. Thus, except with probability $\epsilon/2$, we have $\epsilon/(4(k+1)\lambda) \leq Y_i \leq 4(k+1)/(\epsilon\lambda)$ for $i = 1, 2, \ldots, k+1$. Let $g = 4(k+1)^2/(\epsilon\lambda) + (k+1)$. This ensures that, except with probability $\epsilon/2$, after g time steps, at least $(k+1)$ time steps where the mutator outputs mutations of type "adding" have occurred, and the first k of these occurrences are all separated by at least $\epsilon/(4(k+1)\lambda)$ time steps of other types of mutations.

Also, let $m = p^{-1}\ln(4g/\epsilon)$ and let $s = (200kG^2B^2/\alpha^2)\ln(4mg/\epsilon)$ be the *size function*. These values ensure that for g generations, except with probability $\epsilon/2$, the mutator always produces a mutation that had probability at least p of being produced (conditioned on the type of mutations output by the mutator at that time step), and that for all the representations concerned, $|\hat{\mathrm{L}}_{f,D}(w) - \mathrm{L}_{f,D}(w)| \leq \tau$, where $\tau = \alpha/5$. Thus, allowing the process to fail with probability ϵ, we assume that none of the undesirable events have occurred.

We will show that the steps with mutations other than "adding" are sufficient to ensure that evolution reaches the (almost) best possible target with the variables available to it. In particular, if the set of available variables is S, the representation w reached by evolution will satisfy $\|f^S - w\|^2 \leq \epsilon/(2k^2)$. For now, suppose that this is the case.

We claim by induction that evolution never adds a "wrong" variable, *i.e.*, one that is not present in the target function f. The base case is trivially true, since the starting representation is 0. Now suppose, just before a "heads" step, the representation is w, such that $S = \mathsf{NZ}(w)$ and $\|f^S - w\| \leq \epsilon/(2k^2)$. The current step is assumed to be a "heads" step, thus the mutator has produced mutations by adding a new variable. Then, using Claim 4, we know that there is a mutation w' in $\mathsf{Neigh}(w, \epsilon)$ such that $\mathrm{L}_{f,D}(w') < \mathrm{L}_{f,D}(w) - \epsilon/(4k^2)$ (obtained by adding a correct variable). Since $\alpha < \epsilon/(4k^2)$ and $\tau = \alpha/5$, it must be the case that $\hat{\mathrm{L}}_{f,D}(w') \leq \hat{\mathrm{L}}_{f,D}(w) - 3\alpha/5$. This ensures that the set Best, for selection rule $\mathsf{Opt\text{-}Sel}$ is not empty. Furthermore, we claim that no mutation that adds an *irrelevant* variable can be in Best. Suppose w'' is a mutation that adds an *irrelevant* variable; according to Claim 5, $\mathrm{L}_{f,D}(w'') > \mathrm{L}_{f,D}(w') + \alpha$, and hence $\hat{\mathrm{L}}_{f,D}(w'') > \hat{\mathrm{L}}_{f,D}(w') + t$. This ensures that every representation in Best corresponds to a mutation that adds some *relevant* variable. Thus, the evolutionary algorithm never adds any *irrelevant* variable.

Finally, note that during a "tails" step (when the mutator produces mutations of types other than "adding"), as long as $\|f^S - w\|^2 \geq \epsilon/(4k^2)$, there exists a mutation that reduces the expected loss by at least $\epsilon^2/(192k^4) = \alpha$. This implies that the set Best is non-empty and for the values of tolerance $t = 3\alpha/5$ and $\tau = \alpha/5$, any mutation from the set Best reduces the expected loss by at least $\alpha/5$. (This argument is identical to the one in Theorem 1.) Since the maximum loss is at most $4kB^2G^2$ for the class of distributions and a representation w from the set R; in at most $20kB^2G^2/\alpha$ steps, a representation satisfying $\mathrm{L}_{f,D}(w) \leq \epsilon/(4k^2)$ must be reached. Note that once such a representation is reached, it is ensured that the loss does not increase substantially, since with probability at least $1/2$, the mutator outputs the same representation. Hence, it is guaranteed that there is always a neutral mutation. Thus, before the next "heads" step, it must be the case that $\|f^S - w\|^2 \leq \epsilon/2k^2$. If λ is set to $\epsilon\alpha/(80k(k+1)B^2G^2)$, the evolutionary algorithm using the selection rule $\mathsf{Opt\text{-}Sel}$ succeeds.

It is readily verified that the values of g and s satisfy the claims in the statement of the theorem. □

4 Conclusion and Future Work

In this work, we provided simple evolutionary mechanisms for evolving sparse linear functions, under a large class of distributions. These evolutionary algorithms have the desirable properties that the representations used are themselves sparse linear functions, and that they are attribute-efficient in the sense that the number of generations required for evolution to succeed is independent of the total number of attributes.

Strong negative results are known for distribution-independent evolvability of boolean functions, *e.g.*, even the class of conjunctions is not evolvable [11]. However, along the lines of this work, it is interesting to study whether under restricted classes of distributions, evolution is possible for simple concept classes, using representations of low complexity. Currently, even under (biased) product distributions, no evolutionary mechanism is known for the class of disjunctions, except via Feldman's general reduction from CSQ algorithms. Even if the queries made by the CSQ algorithm are simple, Feldman's reduction uses intermediate representations that randomly combine queries made by the algorithm, making the representations quite complex.

A natural extension of our current results would be to study fixed-degree sparse polynomials. The suitable class of boolean functions to study is low-weight threshold functions, which includes disjunctions and conjunctions. The class of smooth bounded distributions may be an appropriate starting place for studying evolvability of these classes. For example, is the class of low-weight threshold functions evolvable under smooth distributions, or at least log-concave distributions?

Acknowledgments

We would like to thank Leslie Valiant for helpful discussions and comments on an earlier version of this paper. We are grateful to Frank Solomon for discussing biological aspects related to this work.

References

[1] Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, Boca Raton, FL, 2006.

[2] Francis R. Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

[3] S. Balaji, Madan M. Babu, Lakshminarayan M. Iyer, Nicholas M. Luscombe, and L. Aravind. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *Journal of Molecular Biology*, 360(1):213–227, 2006.

[4] Alfred M. Bruckstein, David L. Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.

[5] E. J. Candes and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[6] Cuthbert Daniel and Fred S. Wood. *Fitting Equations to Data: Computer Analysis of Multifactor Data*. John Wiley & Sons, Inc., New York, NY, USA, 2nd edition, 1999.

[7] David L. Donoho, Michael Elad, and Vladimir N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

[8] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[9] Vitaly Feldman. Evolution from learning algorithms. In *Proceedings of the Symposium on the Theory of Computation (STOC)*, 2008.

[10] Vitaly Feldman. Robustness of evolvability. In *Proceedings of the Conference on Learning Theory (COLT)*, 2009.

[11] Vitaly Feldman. Distribution-independent evolution of linear threshold functions. In *Proceedings of the Conference on Learning Theory (COLT)*, 2011.

[12] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–C52, 1999.

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer New York Inc., New York, NY, USA, 2001.

[14] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[15] Adam Tauman Kalai, Alex Samorodnitsky, and Shang-Hua Teng. Learning and smoothed analysis. In *Proceedings of IEEE Conference on the Foundations of Computer Science (FOCS)*, 2009.

[16] Varun Kanade. *Computational Questions in Evolution*. PhD thesis, Harvard University, 2012.

[17] Varun Kanade, Jennifer Wortman Vaughan, and Leslie G. Valiant. Evolution with drifting targets. In *Proceedings of the Conference on Learning Theory (COLT)*, 2010.

[18] Daniel M. Kane, Adam Klivans, and Raghu Meka. Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In *Proceedings of the Conference on Learning Theory (COLT)*, 2013.

[19] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, 1998.

[20] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach. Learn.*, 2(4):285–318, 1988.

[21] Nicholas M. Luscombe, M. Madan Babu, Haiyuan Yu, Michael Snyder, Sarah A. Teichmann, and Mark Gerstein. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, (7006):308–312.

[22] Stphane Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition, 2008.

[23] Douglas B. Murray, Ken Haynes, and Masaru Tomita. Redox regulation in respiring *Saccharomyces cerevisiae*. *Biochimica et Biophysica Acta*, 1810(10):945–958, 2011.

[24] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

[25] Bruno A. Olshausen and David J. Fieldt. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

[26] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3), 2004.

[27] Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[28] Leslie G. Valiant. Evolvability. *Journal of the ACM*, 56(1):3:1–3:21, 2009.

[29] Paul Valiant. Evolvability of real-valued functions. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2012.

[30] Andreas Wagner and David A. Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 268(1478):1803–1810, 2001.

[31] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

A Omitted Proofs

A.1 Proofs from Section 2.2

Proof of Lemma 1. Note that for any $x \sim D$, we can write $x = \tilde{x} + \eta$, where $\tilde{x}$ is drawn from some smooth bounded distribution, and η is drawn from the uniform distribution over $[-\sqrt{3}\Delta, \sqrt{3}\Delta]^n$. Note that η and $\tilde{x}$ are independent, and all components of η are independent. First, we observe that $\mathbb{E}[x_i^2] = \mathbb{E}[(\tilde{x}_i + \eta_i)^2] \geq \mathbb{E}[\eta_i^2] = \Delta^2$. Now, consider the following:

$$\begin{aligned}
\langle w, w \rangle &= \mathbb{E}\left[\left(\sum_{i=1}^{n} w_i x_i\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} w_i(\tilde{x}_i + \eta_i)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} w_i \tilde{x}_i\right)^2\right] + 2\mathbb{E}\left[\left(\sum_{i=1}^{n} w_i \tilde{x}_i\right)\left(\sum_{i=1}^{n} w_i \eta_i\right)\right] + \mathbb{E}\left[\left(\sum_{i=1}^{n} w_i \eta_i\right)^2\right] \\
&\geq \sum_{i=1}^{n} w_i^2 \mathbb{E}[\eta_i^2] \qquad \text{Since } \eta_i \text{ are all independent} \\
&\geq \sum_{i=1}^{n} w_i^2 \Delta^2 \qquad \text{Since } \mathbb{E}[\eta_i^2] = \Delta^2 \text{ by definition} \\
&= \sum_{i \in \mathsf{NZ}(w)} w_i^2 \Delta^2
\end{aligned}$$

The conclusions of the Lemma follow easily by looking at the above expression. □

A.2 Proofs from Section 3.1

Proof of Claim 1. We show that in this case, a "scaling" mutation achieves the desired result. Restricted to the direction w, the best approximation to f^S is $\frac{\langle f^S, w \rangle}{\|w\|^2} w$. We have that

$$\left\| \frac{\langle f^S, w \rangle}{\|w\|^2} w \right\| \leq \|f^S\| \leq \frac{\|w\|}{2}$$

Hence, if $\langle f^S, w \rangle > 0$, for $\gamma \in [1/4, 3/4]$ (and similarly if $\langle f^S, w \rangle < 0$ for $\gamma \in [-3/4, -1/4]$), we have that

$$\begin{aligned}
\|f^S - \gamma w\|^2 &= \|f^S - w\|^2 + 2(1-\gamma)\langle f^S - w, w \rangle + (1-\gamma)^2 \|w\|^2 \\
&\leq \|w - f^S\|^2 - (1-\gamma)\|w\|^2 + (1-\gamma)^2 \|w\|^2 \\
&= \|w - f^S\|^2 - (\gamma - \gamma^2)\|w\|^2
\end{aligned}$$

Finally, by observing that for $\gamma \in [1/4, 3/4]$, $\gamma - \gamma^2 \geq 3/16$ and that by the triangle inequality, $\|w\| \geq (2/3)\|f^S - w\|$ when $\|w\| \geq 2\|f^S\|$, we obtain

$$\|f^S - \gamma w\|^2 \leq \|f^S - w\|^2 - \frac{1}{12}\|f^S - w\|^2$$

We note that $\mathrm{L}_{f,D}(w') = \|f - f^S\|^2 + \|f^S - \gamma w\|^2$ and $\mathrm{L}_{f,D}(w) = \|f - f^S\|^2 + \|f^S - w\|^2$. An appropriate value of γ is chosen with probability at least $1/4$, and combined with the probability of choosing a scaling mutation we get the desired result. □

Proof of Claim 2. Here, we appeal to a mutation that adjusts the relative weights of the variables within the set $S = \mathsf{NZ}(w)$. Consider the vector $f^S - w$, and note that $\mathsf{NZ}(f^S - w) \subseteq S$. Let $r^S = f^S - w$ denote the residual, which lies in the space spanned by S. Now consider

$$\|r_S\|^2 = \langle r^S, r^S \rangle = \sum_{i \in S} r_i^S \langle e^i, r^S \rangle$$

Here, e^i is the unit vector representing the linear function $x \mapsto e^i \cdot x = x_i$. Therefore, there must exist an i for which the following is true:

$$r_i^S \langle e^i, r^S \rangle \geq \frac{\|r^S\|^2}{|S|}$$

We appeal to Lemma 1 (part 1), which implies that $|r_i^S| \leq \sqrt{\langle r^S, r^S \rangle / \Delta^2} = \|r^S\|/\Delta$, to conclude that

$$|\langle e^i, r^S \rangle| \geq \frac{\|r^S\|\Delta}{|S|}$$

Let $\beta = \frac{\langle e^i, r^S \rangle}{\|e^i\|^2}$ and suppose $w' = w + \gamma e^i$ for $\gamma \in [\beta - |\beta|/2, \beta + |\beta|/2]$. We then have

$$\|f^S - (w + \gamma e^i)\|^2 = \|f^S - w\|^2 - 2\gamma \langle f^S - w, e^i \rangle + \gamma^2 \|e^i\|^2$$

Recall that $f^S - w = r^S$ and note that $\langle e^i, r^S \rangle$ and γ have the same sign. This, combined with the above equation, gives

$$\|f^S - (w + \gamma e^i)\|^2 = \|f^S - w\|^2 - (2|\gamma||\beta| - \gamma^2)\|e^i\|^2$$

Finally, note that for $|\gamma| \in [|\beta|/2, 3|\beta|/2]$, $-2|\gamma||\beta| + \gamma^2 \leq -3|\beta|^2/4$, hence, the above equation and the fact that $\|e^i\| \leq 1$ together yield

$$\|f^S - (w + \gamma e^i)\|^2 \leq \|f^S - w\|^2 - \frac{3}{4}\beta^2 \|e^i\|^2 \leq \|f^S - w\|^2 - \frac{3}{4}\frac{\|f^S - w\|^2 \Delta^2}{|S|^2}$$

Note that a suitable mutation $w' = w + \gamma e^i$ is obtained with probability at least $|\beta|/(6KB)$ ($1/3$ for choosing the right type of mutation, $1/K$ for the correct choice of variable, and $|\beta|/(2B)$ for the choosing the correct value of w_i). Also note that $|\beta| \geq \Delta \|f^S - w\|/|S| \geq \Delta \|f^S - w\|/K$. For this to be a valid mutation, we also need to verify the fact that $w_i + \gamma \in [-B, B]$, which is ensured by our choice of B. To see this, note that $\|w\| \leq 2\|f^S\| \leq 2\|f\|$ (the last part is because f^S is a projection of f onto a lower dimensional space). Thus, by Lemma 1 (part 1), $w_i \leq 2\|f\|/\Delta$. Also, $|\beta| = |\langle e^i, r^S \rangle|/\|e^i\|^2 \leq \|r^S\|/\|e^i\| = \|f^S - w\|/\mathbb{E}[x_i^2] \leq 3\|f\|/\Delta$. Thus, if $B > 13\|f\|/(2\Delta)$, then $|w_i + \gamma| < B$ and the mutation will be a valid one. Note that the maximum value of $\|f\|$ for $f \in \mathsf{Lin}_{l,u}^k$ is uk. Thus, our choice of $B = 10uk/\Delta$ is sufficient. This completes the proof of Claim 2. □

Proof of Claim 3. Finally, we show that if $\|f^S - w\|$ is very small, but $\mathsf{NZ}(f) \not\subseteq S$, then it must be the case that a "swapping" or "adding" mutation is beneficial. We focus on the swapping case, *i.e.*, when $|S| = K$; the adding step is a special case of this. First, we observe that if there exists $i \in \mathsf{NZ}(f)$ such that $i \notin S$, then by using Lemma 1 (part 1), it must be the case that $\|f - w\|^2 \geq (f_i - w_i)^2\Delta^2 = f_i^2\Delta^2 \geq l^2\Delta^2$. Let $r = f - w$ denote the residual. Then, consider the following:

$$\langle r, r\rangle = \sum_{i \in \mathsf{NZ}(f)\setminus S} r_i\langle e^i, r\rangle + \sum_{i \in S} r_i\langle e^i, r\rangle$$

Note that for all $i \in S$, $\langle e^i, f - f^S\rangle = 0$, since f^S is the projection of f onto the space spanned by the variables in S. Hence, if $r^S = f^S - w$, the residual within the space spanned by S, then $r = f - f^S + r^S$. Thus, we have $\langle e^i, r\rangle = \langle e^i, r^S\rangle \leq \|r^S\|$. Using this we get,

$$\langle r, r\rangle \leq \sum_{i \in \mathsf{NZ}(f)\setminus S} r_i\langle e^i, r\rangle + \|r^S\| \sum_{i \in S} |r_i|$$

Now, even by a very crude estimate, $|r_i| = |f_i - w_i| \leq 2B$, and hence by the condition in the statement of Claim 3, that $\|r^S\| = \|f^S - w\| \leq l^2\Delta^2/(4KB)$, together with the previous observation that $\langle r, r\rangle = \|r\|^2 = \|f - w\|^2 \geq l^2\Delta^2$, we have that

$$\frac{1}{2}\langle r, r\rangle \leq \sum_{i \in \mathsf{NZ}(f)\setminus S} r_i\langle e^i, r\rangle$$

We now appeal to Lemma 1 (part 1), which shows that $|r_i| \leq \|r\|/\Delta$, and conclude that there exists an i for which

$$|\langle e^i, r\rangle| \geq \frac{\|r\|\Delta}{2k}$$

The crucial observation is that $|\mathsf{NZ}(f)| \leq k < K$. Let $\beta = \frac{\langle r, e^i\rangle}{\|e^i\|^2}$. Finally, Lemma 1 (part 2) implies that there exists an i' for which $w_{i'}^2 \leq \|w\|^2/K$. We consider the mutation, $w' = w + \gamma e^i - w_{i'}e^{i'}$ for $\gamma \in [\beta - |\beta|/2, \beta + |\beta|/2]$. Then, we have

$$\|f - (w + \gamma e^i - w_{i'}e^{i'})\|^2 = \|f - (w + \gamma e^i)\|^2 - 2w_{i'}\langle f - (w + \gamma e^i), e^{i'}\rangle + w_{i'}^2\|e^{i'}\|^2 \quad (2)$$

We bound the first term on the right hand side of the above expression and then the latter two.

$$\begin{aligned}\|f - (w + \gamma e^i)\|^2 &= \|r\|^2 - 2\gamma\langle r, e^i\rangle + \gamma^2\|e^i\|^2 \\ &= \|r\|^2 - (2\gamma\beta - \gamma^2)\|e^i\|^2\end{aligned}$$

As in the proof of Claim 2, for $|\gamma| \in [|\beta|/2, 3|\beta|/2]$, we have that $-2\gamma\beta + \gamma^2 \leq -3\beta^2/4$. Hence,

$$\|f - (w + \gamma e^i)\|^2 \leq \|r\|^2 - \frac{3}{4}\beta^2 \quad (3)$$

To bound the remaining two terms in (2), recall that $f - w = f - f^S + r^S$ and that $\langle f - f^S, e^{i'}\rangle = 0$ (since $i' \in S$). Thus, we get that

$$-2w_{i'}\langle f - (w + \gamma e^i), e^{i'}\rangle + w_{i'}^2\|e^{i'}\|^2 \leq 2|w_{i'}||\langle r^S + \gamma e^i, e^{i'}\rangle| + w_{i'}^2\|e^{i'}\|^2$$

Using the fact that $\|r^S\| \leq \gamma$, $|w_{i'}| < \gamma$ (which can be verified by the setting of K below), $|\langle e^i, e^{i'}\rangle| \leq 1$ and $\|e^{i'}\| \leq 1$, we obtain

$$-2w_{i'}\langle f - (w + \gamma e^i), e^{i'}\rangle + w_{i'}^2\|e^{i'}\|^2 \leq 6|w_{i'}||\gamma| \tag{4}$$

Recall that $|w_{i'}| \leq \|w\|/\sqrt{K}$. Also $\|w\| \leq \|f^S\| + \|r^S\| \leq 2\|f^S\| \leq 2\|f\| \leq 2uk$, where we have used the fact that $\|r^S\|$ is small. Combining (2), (3), (4), the fact that $|w_i| \leq 2uk/\sqrt{K}$ and $|\gamma| \leq 3|\beta|/2$, we get

$$\|f - (w + \gamma e^i - w_{i'}e^{i'})\|^2 \leq \|r\|^2 - \frac{3}{4}\beta^2 + 18|\beta|uk/\sqrt{K}$$

Finally, we note that when $K > 5184(k/\Delta)^4(u/l)^2$, the above equation ensures that the expected loss drops by at least $\beta^2/4$. The probability of choosing a swapping operations is $1/3$, of subsequently choosing the correct pair is at least $1/(nK)$, and subsequently choosing the correct value of w_i is at least $|\beta|/(2B)$. A simple calculation proves the statement of the claim. □

A.3 Proofs from Section 3.2

Proof of Claim 4. Since $S \subseteq \mathsf{NZ}(f)$, the residual $r = f - w$ is such that $\mathsf{NZ}(r) \subseteq \mathsf{NZ}(f)$. Consider $i \in \mathsf{NZ}(r)$ that maximizes $|r_i| \cdot \|e^i\|$. Then, we have:

$$\begin{aligned}\left|\frac{\langle e^i, r\rangle}{\|e^i\|}\right| &\geq \frac{|r_i|\langle e^i, e^i\rangle}{\|e^i\|} - \sum_{j \in \mathsf{NZ}(r), j \neq i} \frac{|r_j|\langle e^i, e^j\rangle}{\|e^i\|} \\ &\geq |r_i| \cdot \|e^i\| - \frac{1}{2k} \sum_{j \in \mathsf{NZ}(r), j \neq i} |r_j| \cdot \|e^j\| \\ &\geq |r_i| \cdot \|e^i\| \cdot \frac{k+1}{2k},\end{aligned} \tag{5}$$

where in the last two steps we used the fact that $\mathrm{corr}(x_i, x_j) = \langle e^i, e^j\rangle/(\|e^i\|\|e^j\|) \leq 1/(2k)$ and that $|\mathsf{NZ}(r) \setminus \{i\}| \leq k - 1$. On the other hand, for any $i' \notin \mathsf{NZ}(r)$, we have

$$\begin{aligned}\left|\frac{\langle e^{i'}, r\rangle}{\|e^{i'}\|}\right| &\leq \sum_{j \in \mathsf{NZ}(r)} \frac{|r_j\langle e^{i'}, e^j\rangle|}{\|e^{i'}\|} \\ &\leq \frac{1}{2k} \sum_{j \in \mathsf{NZ}(r)} |r_j| \cdot \|e^j\| \leq \frac{1}{2}|r_i| \cdot \|e^i\|\end{aligned} \tag{6}$$

Here, again in the last two steps, we have used the fact that $\mathrm{corr}(x_{i'}, x_j) \leq 1/(2k)$ and that $|r_i| \cdot \|e^i\|$ is the largest such term.

First, we claim that if $\|r\|^2 \geq \epsilon$, then the i that maximized $|r_i| \cdot \|e^i\|$ must be from the set $\mathsf{NZ}(f) \setminus S$. (Note that $\|r\|^2 = \|f - w\|^2 = \mathrm{L}_{f,D}(w)$, so if $\|r\|^2 \leq \epsilon$, evolution has reached its goal.) By the triangle inequality, $\sum_{i \in \mathsf{NZ}(r)} |r_i| \cdot \|e^i\| \geq \|r\| \geq \sqrt{\epsilon}$. Hence, it must be the case that $|r_i| \cdot \|e^i\| \geq \sqrt{\epsilon}/k$. For contradiction, assume that $i \in S$. Then, since f^S is the projection of f in the space spanned by S, we have $\langle e^i, r\rangle = \langle e^i, f^S - w\rangle$, since $r = f - f^S + f^S - w$ and $\langle e^i, f - f^S\rangle = 0$. But, by the assumption of the claim, $|\langle e^i, r\rangle| \leq \|e^i\| \cdot \|f^S - w\| \leq \|e^i\| \cdot \sqrt{\epsilon}/(2k)$, and by (5), we know that $|\langle e^i, r\rangle| \geq \|e^i\| \cdot (|r_i| \cdot \|e^i\|) \cdot (k+1)/(2k) > \|e^i\| \cdot \|r\|/(2k) \geq \|e^i\| \cdot \sqrt{\epsilon}/(2k)$. Thus, it cannot be the case that $i \in S$.

Let $w' = w + \gamma e^i$. Then,

$$\begin{aligned}\|f - (w+\gamma e^i)\|^2 - \|f-w\|^2 &= -2\gamma\langle f - w, e^i\rangle + \gamma^2\|e^i\|^2 \\ &\leq -\|e^i\|^2\left(|\gamma|\cdot|r_i|\cdot\frac{k+1}{k} - |\gamma|^2\right)\end{aligned}$$

Now suppose γ satisfies $1-\delta \leq (2|\gamma|k)/(|r_i|(k+1)) \leq 1+\delta$, then using the fact that the quadratic function on the RHS is maximized at $|\gamma| = (k+1)|r_i|/(2k)$, we have

$$\|f - (w+\gamma e^i)\|^2 - \|f-w\|^2 \leq -\|e^i\|^2 \cdot \frac{r_i^2}{4}\cdot\frac{(k+1)^2}{k^2}\cdot(1-\delta^2) \tag{7}$$

Note that for any $i' \notin S$, the "best" representation of the form $w+\beta e^{i'}$ is when $\beta = \langle e^{i'}, r\rangle/\|e^{i'}\|^2$, and the corresponding reduction in squared loss is $(\langle e^{i'}, r\rangle)^2/\|e^{i'}\|^2$. Thus, for any $i' \neq i$, we have

$$\begin{aligned}\|f - (w+\beta e^{i'})\|^2 - \|f-w\|^2 &\leq -\frac{\langle e^{i'}, r\rangle^2}{\|e^{i'}\|^2} \\ &\leq -\frac{r_i^2}{4}\|e^i\|^2 \qquad \text{Using (6)}\end{aligned}$$

Setting $\delta = \sqrt{1/(k+1)}$ completes the proof of the claim. To see that $b - a \geq \sqrt{(k+1)\epsilon}/k^2$, notice that any γ, such that $|\gamma| \in [(1-\delta)((k+1)/(2k))|r_i|, (1-\delta)((k+1)/(2k))|r_i|]$, achieves the claimed reduction in squared loss. Since $|r_i|\cdot\|e^i\| \geq \sqrt{\epsilon}/k$, we have that $|r_i| \geq \sqrt{\epsilon}/k$. Hence, $b - a \geq 2\delta((k+1)/2k)(\sqrt{\epsilon}/k) \geq \sqrt{(k+1)\epsilon}/k^2$, for $\delta = \sqrt{1/(k+1)}$. □

Proof of Claim 5. The proof follows along the lines of the proofs of Claims 1 and 2. First, suppose that $\|w\| \geq 2\|f^S\|$. In this case, we claim that for $\gamma \in [1/2, 3/4]$, the mutation γw reduces the squared loss by at least $\|f^S - w\|^2/12$. This analysis is completely identical to that in Claim 1 and hence is omitted. The only difference is that the probability that such a mutation is selected is $1/16$, conditioned on the event that the mutator chooses mutations that don't add an extra variable.

Next, we assume that $\|w\| \leq 2\|f^S\|$. Let $r^S = f^S - w$. Now, as in the proof of Claim 4, consider $i \in \mathsf{NZ}(r^S)$ (recall that $\mathsf{NZ}(r^S) = S$) that maximizes $|r_i^S|\cdot\|e^i\|$. Then, the following is true:

$$\begin{aligned}\left|\frac{\langle e^i, r^S\rangle}{\|e^i\|}\right| &\geq |r_i^S|\frac{\langle e^i, e^i\rangle}{\|e^i\|} - \sum_{j\in S, j\neq i}|r_j^S|\frac{\langle e^i, e^j\rangle}{\|e^i\|} \\ &\geq |r_i^S|\cdot\|e^i\| - \frac{1}{2k}\sum_{j\in S, j\neq i}|r_j^S|\cdot\|e^j\| \\ &\geq |r_i^S|\cdot\|e^i\|\cdot\frac{k+1}{2k} \end{aligned}\tag{8}$$

where in the last two steps we used the fact that $\mathrm{corr}(x_i, x_j) \leq 1/(2k)$ and that $|r_j^S|\cdot\|e^j\|$ is maximized for $j = i$, and the fact that $|S| \leq k$. Also, by the triangle inequality, we know that $\sum_{j\in S}|r_j^S|\cdot\|e^j\| \geq \|r^S\|$; hence, by definition of i, we have that $|r_i^S|\cdot\|e^i\| \geq \|r^S\|/k$. Now, let $\beta = \langle e^i, r^S\rangle/\|e^i\|^2$, and for $\gamma \in [\beta - |\beta|/2, \beta + |\beta|/2]$, consider the mutation $w + \gamma e^i$. We have,

$$\begin{aligned}\|f^S - (w+\gamma e^i)\|^2 - \|f^S - w\|^2 &= -2\gamma\langle e^i, r^S\rangle + \gamma^2\|e^i\|^2 \\ &= -\|e^i\|^2(2|\gamma||\beta| - |\gamma|^2) \\ &\leq -\frac{3}{4}\|e^i\|^2\beta^2 \qquad \text{For } \gamma \in [\beta - |\beta|/2, \beta + |\beta|/2] \\ &\leq -\frac{3}{16}\frac{\|r^S\|^2}{k^2} \qquad \text{Using (8) and the defn. of } \beta\end{aligned}$$

In order for $w + \gamma e^i$ to be a valid mutation, we need to check that $|w_i| + 3|\beta|/2 < B$. To see this, observe the following:

$$\begin{aligned}
\|w\|^2 &\geq \sum_{j \in \mathsf{NZ}(w)} w_j^2 \|e^j\|^2 - \sum_{j_1 < j_2; j_1, j_2 \in \mathsf{NZ}(w)} 2|w_{j_1} w_{j_2}||\langle e^{j_1}, e^{j_2} \rangle| \\
&\geq \sum_{j \in \mathsf{NZ}(w)} w_j^2 \|e^j\|^2 - \frac{1}{k} \sum_{j_1 < j_2} |w_{j_1} w_{j_2}| \|e^{j_1}\| \|e^{j_2}\| \\
&\geq \frac{1}{2} \sum_{j \in \mathsf{NZ}(w)} w_j^2 \|e^j\|^2 + \frac{1}{2k} \sum_{j_1 < j_2} (|w_{j_1}| \|e^{j_1}\| - |w_{j_2}| \|e^{j_2}\|)^2 \\
&\geq \frac{1}{2} \sum_{j \in \mathsf{NZ}(w)} w_j^2 \|e^j\|^2
\end{aligned}$$

Hence, by a fairly loose analysis, $|w_i| \leq 2\|w\|/\Delta \leq 4\|f\|/\Delta \leq 4uk/\Delta$ (since $\|e^i\| \geq \Delta$ for the class of distributions defined in Defn. 3). Also $|\beta| = |\langle e^i, r^S \rangle|/\|e^i\|^2 \leq \|r^S\|/\|e^i\| \leq 3\|f\|/\Delta$ (since $\|r^S\| \leq \|f^S\| + \|w\| \leq 3\|f^S\| \leq 3\|f\|$). It's easy to see that $|w_i| + (3/2)|\beta| \leq B$, for $B = 10uk/\Delta$.

Finally, note that conditioned on the mutator choosing a mutation that doesn't add new variables, the probability of choosing such a mutation is at least $|\beta|/(8Bk)$ (1/4 for choosing a mutation of type "adjusting", $1/k$ for choosing the appropriate variable to adjust and $|\beta|/(2B)$ for choosing the correct value). Combining the claims for mutations of the "scaling" and "adjusting" types and taking the appropriate minimum values proves the statement of the claim. □