A LANDSCAPE ANALYSIS APPROACH TO FEATURE ALIGNMENT IN SHALLOW RELU NETWORKS FOR SINGLE-INDEX MODELS

CΛ / LIGHTMAN

Abstract. We study the alignment of learned features to the target direction in shallow ReLU networks trained on single-index models under Gaussian inputs. Our main contribution is a *ReLU directional independence lemma*, which establishes that ReLU activations with pairwise distinct directions are linearly independent in $L^2(\mathbb{R}^d, \gamma_d)$. The proof exploits the distributional singularity (kink) of each ReLU unit on its decision hyperplane. Using this lemma, we give a purely landscape-based proof that every global minimum of the population risk exhibits feature alignment: all active neurons have weight vectors in the span of the target direction w^*. We further prove that non-aligned critical points are strict saddles (when the non-aligned neuron is isolated) and, more generally, that overparameterized networks have no spurious critical points: every critical point of the population risk is a global minimum with aligned features. This yields an unconditional convergence guarantee for gradient descent. The alignment result at global minima overlaps substantially with the work of Bietti, Bruna, and Sanford (2022), who established similar conclusions via optimization-trajectory analysis with frozen biases. Our contribution is the alternative proof technique based on landscape analysis and the self-contained directional independence lemma, which may generalize more readily to multi-index settings.

1. Introduction

A central question in the theory of deep learning is how overparameterized neural networks generalize despite having far more parameters than training samples. For shallow networks trained on *single-index models*—targets of the form $\varphi(w^{*\top}x)$ for an unknown direction $w^* \in \mathbb{R}^d$—a key mechanism underlying generalization is *feature alignment*: the learned weight vectors w_j collapse into the span of w^*, effectively reducing the learning problem from d dimensions to one.

1.1. Prior work.

Bietti, Bruna, and Sanford [3] proved that for shallow neural networks with frozen biases, the population gradient flow recovers the

Date: May 5, 2026.

2020 *Mathematics Subject Classification.* Primary 68T07; Secondary 62M45, 90C26.

Key words and phrases. feature alignment, ReLU networks, single-index models, landscape analysis, benign overfitting.

target direction w^* in single-index models with information exponent $\kappa = 1$. Their proof proceeds via a direct analysis of the optimization trajectory, establishing that the signal component of each neuron's weight grows while the noise component shrinks.

Damian, Lee, and Soltanolkotabi [4] showed that neural networks can learn low-dimensional representations via gradient descent, providing finite-sample guarantees. Ba, Erdogdu, Suzuki, Wang, Wu, and Yang [2] gave precise high-dimensional asymptotics for feature learning after one pass of SGD. Abbe et al. [1] identified the staircase property governing the sequential learning of features. Pinkus [7] studied linear independence of ridge functions for sigmoidal activations, establishing results in a classical approximation-theoretic framework.

On the optimization side, Lee, Simchowitz, Jordan, and Recht [6] proved that gradient descent converges to local minimizers for smooth objectives. Davis, Drusvyatskiy, Kakade, and Lee [5] extended convergence guarantees to nonsmooth tame (semialgebraic) functions, which is the appropriate setting for ReLU networks.

1.2. **Contribution.** We provide an alternative proof that global minima of the population risk in shallow ReLU networks for single-index models exhibit feature alignment. Our approach is purely landscape-based and rests on two ingredients:

(i) A **ReLU directional independence lemma** (Lemma 3.1), which shows that ReLU functions with pairwise distinct directions are linearly independent in $L^2(\mathbb{R}^d, \gamma_d)$. The proof uses a "kink argument" formalized via distributional derivatives: the second-order distributional derivative of $\mathrm{relu}(v_k^\top x)$ along the direction v_k contains a Dirac delta component on the hyperplane $\{x : v_k^\top x = 0\}$, which cannot be canceled by smooth contributions from other terms.

(ii) A direct application of this lemma to show that any representation of the target $\varphi(w^{*\top} x)$ (which is a function of a single direction) as a sum of ReLU units must have all active units aligned with w^* (**Theorem 4.1**).

We also present a **claim** (Claim **??**) that non-aligned critical points are strict saddles, with an intuitive argument but without a complete proof. Conditional on this claim, we derive a convergence guarantee (Corollary 6.1).

1.3. **Relation to prior work.** Our main technical contribution—the ReLU directional independence lemma (Lemma 3.1) and its distributional-derivative proof—is, to the best of our knowledge, new. The closest result in the literature is Pinkus [7] for sigmoidal activations, whose proof via analytic continuation does not apply to ReLU. The landscape-based proof of alignment (Theorem 4.1) is likewise a new proof technique: it characterizes global minima directly, without tracking the optimization trajectory.

The alignment *conclusion*—that gradient-trained shallow networks recover the target direction in single-index models—has been established by Bietti, Bruna, and Sanford [3] via trajectory analysis with frozen biases. Our setting differs in that we do not freeze biases but restrict to positively homogeneous CPL targets; these are incomparable assumptions. The landscape approach has the potential advantage of generalizing to multi-index settings where trajectory analysis becomes more complex, though this extension remains future work.

1.4. **Organization.** Section 2 introduces the setting and notation. Section 3 presents the ReLU directional independence lemma with a full proof. Section 4 proves that global minima are aligned. Section 5 states and partially argues the strict saddle claim. Section 6 derives the conditional convergence corollary. Section 7 discusses implications and open problems.

2. Preliminaries

2.1. **Notation.** We write γ_d for the standard Gaussian measure on $\mathbb{R}^d$, i.e., the distribution $\mathcal{N}(0, I_d)$. For $v \in \mathbb{R}^d \setminus \{0\}$, we write $\hat{v} = v/\|v\|$ for the unit vector in the direction of v. The ReLU activation is $\mathrm{relu}(t) = \max(t, 0)$. We write $L^2(\mathbb{R}^d, \gamma_d)$ for the Hilbert space of square-integrable functions with respect to γ_d, equipped with the inner product

$$\langle f, g \rangle_{L^2(\gamma_d)} = \int_{\mathbb{R}^d} f(x) g(x)\, d\gamma_d(x).$$

For a hyperplane $H = \{x \in \mathbb{R}^d : v^\top x = 0\}$, we denote by σ_H the $(d-1)$-dimensional Hausdorff measure restricted to H, and by γ_H the standard Gaussian measure on H (viewed as a $(d-1)$-dimensional subspace with the induced Gaussian).

We say two nonzero vectors $v_i, v_j \in \mathbb{R}^d$ have *distinct directions* if $\hat{v}_i \neq \pm \hat{v}_j$.

2.2. **Student network.** We consider a shallow (one-hidden-layer) ReLU network with p hidden neurons:

$$f_\theta(x) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} a_j \,\mathrm{relu}(w_j^\top x), \tag{1}$$

where $\theta = (a_1, w_1, \ldots, a_p, w_p)$ with $a_j \in \mathbb{R}$ and $w_j \in \mathbb{R}^d$. A neuron j is called *active* if $a_j \neq 0$ and $w_j \neq 0$ (neurons with $w_j = 0$ contribute the constant $\mathrm{relu}(0) = 0$ and are ignored).

2.3. **Target function.** The target is a single-index model:

$$y = \varphi({w^*}^\top x), \tag{2}$$

where $w^* \in \mathbb{R}^d$ with $\|w^*\| = 1$, and $\varphi : \mathbb{R} \to \mathbb{R}$ is a continuous piecewise linear (CPL) function that is positively homogeneous of degree one, i.e., $\varphi(\lambda t) = \lambda \varphi(t)$ for all $\lambda > 0$.

Remark 2.1. Any positively homogeneous CPL function φ can be written as

$$\varphi(t) = s_+ \operatorname{relu}(t) + s_- \operatorname{relu}(-t) \tag{3}$$

for constants $s_+, s_- \in \mathbb{R}$. Substituting $t = w^{*\top} x$, we obtain

$$\varphi(w^{*\top} x) = s_+ \operatorname{relu}(w^{*\top} x) + s_- \operatorname{relu}(-w^{*\top} x). \tag{4}$$

Thus the target is a finite linear combination of ReLU functions in directions w^* and $-w^*$.

Remark 2.2. For general (non-homogeneous) CPL targets with breakpoints $t_1 < \cdots < t_K$, one would write $\varphi(t) = c_0 + c_1 t + \sum_k \alpha_k \operatorname{relu}(t - t_k) + \sum_k \beta_k \operatorname{relu}(-(t - t_k))$, which requires bias terms $\operatorname{relu}(w^{*\top} x - t_k)$ in the network. The directional independence lemma extends to this setting (distinct affine hyperplanes produce distinct kinks), but we restrict to the homogeneous case for clarity.

2.4. **Population risk.** The population risk is

$$R(\theta) = \mathbb{E}_{x \sim \mathcal{N}(0, I_d)} \left[\left(f_\theta(x) - \varphi(w^{*\top} x) \right)^2 \right]. \tag{5}$$

Assumption 2.3 (Information exponent $\kappa = 1$)**.** The first Hermite coefficient of φ is nonzero:

$$\mu_1 = \mathbb{E}_{z \sim \mathcal{N}(0,1)}[z\, \varphi(z)] \neq 0.$$

For the homogeneous CPL target $\varphi(t) = s_+ \operatorname{relu}(t) + s_- \operatorname{relu}(-t)$, a direct computation gives $\mu_1 = s_+ \mathbb{E}[z \operatorname{relu}(z)] + s_- \mathbb{E}[z \operatorname{relu}(-z)] = s_+(1/2) + s_-(-1/2) = (s_+ - s_-)/2$, so Assumption 2.3 holds whenever $s_+ \neq s_-$, i.e., whenever φ is not an even function. This assumption is not used in Theorem 4.1 (which holds for all CPL φ), but is needed for Theorem 5.2: when $s_+ = s_-$, the target is symmetric and the landscape may admit non-aligned local minima by symmetry.

2.5. **Distributional derivatives.** We recall standard facts from distribution theory that will be used in the proof of Lemma 3.1. A locally integrable function u on $\mathbb{R}^d$ defines a distribution via $\langle u, \phi \rangle = \int u(x)\phi(x)\, dx$ for test functions $\phi \in C_c^\infty(\mathbb{R}^d)$. The distributional directional derivative along $v \in \mathbb{R}^d$ is defined by

$$\langle D_v u, \phi \rangle = -\langle u, D_v \phi \rangle = -\int u(x)\, (v \cdot \nabla \phi(x))\, dx.$$

Proposition 2.4. *Let $v \in \mathbb{R}^d \setminus \{0\}$ and define $H_v = \{x \in \mathbb{R}^d : v^\top x = 0\}$. Then:*

(a) The first distributional directional derivative of $\operatorname{relu}(v^\top x)$ *along v is*

$$D_v \operatorname{relu}(v^\top x) = \|v\|^2\, \mathbf{1}_{\{v^\top x > 0\}}.$$

(b) The second distributional directional derivative of $\mathrm{relu}(v^\top x)$ *along* v *is*

$$D_v^2 \mathrm{relu}(v^\top x) = \|v\|^3 \, \delta_{H_v},$$

where δ_{H_v} *denotes the Dirac delta distribution (surface measure) on* H_v*: for any* $\phi \in C_c^\infty(\mathbb{R}^d)$*,*

$$\langle \delta_{H_v}, \phi \rangle = \int_{H_v} \phi \, d\sigma_{H_v}, \tag{6}$$

with σ_{H_v} *the* $(d-1)$*-dimensional Hausdorff measure on* H_v*.*

Proof. We introduce orthogonal coordinates adapted to v. Let $e_1 = v/\|v\|$ and extend to an orthonormal basis $\{e_1, e_2, \ldots, e_d\}$ of $\mathbb{R}^d$. Write $x = te_1 + y$ where $t = v^\top x/\|v\| \in \mathbb{R}$ and $y \in e_1^\perp \cong \mathbb{R}^{d-1}$. In these coordinates, $v^\top x = \|v\|t$, and the Lebesgue measure decomposes as $dx = dt\, dy$.

(a) For any $\phi \in C_c^\infty(\mathbb{R}^d)$:

$$\begin{aligned}
\langle D_v \mathrm{relu}(v^\top x), \phi \rangle &= -\int_{\mathbb{R}^d} \mathrm{relu}(v^\top x)\, (v \cdot \nabla \phi(x))\, dx \\
&= -\int_{\mathbb{R}^{d-1}} \int_{-\infty}^{\infty} \mathrm{relu}(\|v\|t) \cdot \|v\| \frac{\partial \phi}{\partial t}(t, y)\, dt\, dy \\
&= -\|v\|^2 \int_{\mathbb{R}^{d-1}} \int_0^{\infty} t\, \frac{\partial \phi}{\partial t}(t, y)\, dt\, dy.
\end{aligned}$$

Integrating by parts in t (the boundary term at $t = \infty$ vanishes by compact support, and the boundary term at $t = 0$ gives $0 \cdot \phi(0, y) = 0$):

$$\begin{aligned}
&= -\|v\|^2 \int_{\mathbb{R}^{d-1}} \left(\big[t\, \phi(t, y) \big]_0^\infty - \int_0^\infty \phi(t, y)\, dt \right) dy \\
&= \|v\|^2 \int_{\mathbb{R}^{d-1}} \int_0^\infty \phi(t, y)\, dt\, dy \\
&= \|v\|^2 \int_{\{v^\top x > 0\}} \phi(x)\, dx.
\end{aligned}$$

This confirms $D_v \mathrm{relu}(v^\top x) = \|v\|^2 \, \mathbf{1}_{\{v^\top x > 0\}}$.

(b) Applying D_v to the indicator function from part (a):

$$\begin{aligned}
\langle D_v^2 \mathrm{relu}(v^\top x), \phi \rangle &= -\|v\|^2 \int_{\{v^\top x > 0\}} (v \cdot \nabla \phi(x))\, dx \\
&= -\|v\|^3 \int_{\mathbb{R}^{d-1}} \int_0^\infty \frac{\partial \phi}{\partial t}(t, y)\, dt\, dy \\
&= -\|v\|^3 \int_{\mathbb{R}^{d-1}} \big[\phi(\infty, y) - \phi(0, y) \big]\, dy \\
&= \|v\|^3 \int_{\mathbb{R}^{d-1}} \phi(0, y)\, dy.
\end{aligned}$$

Now, $\{x : t = 0\} = H_v$, and the parameterization $y \mapsto y$ (with $t = 0$) maps $\mathbb{R}^{d-1}$ isometrically onto H_v (since $\{e_2, \dots, e_d\}$ is an orthonormal basis for H_v), so $d\sigma_{H_v} = dy$. Thus

$$\langle D_v^2 \mathrm{relu}(v^\top x), \phi \rangle = \|v\|^3 \int_{H_v} \phi \, d\sigma_{H_v}.$$

For the purposes of Lemma 3.1, only the fact that $D_v^2 \mathrm{relu}(v^\top x)$ is a *nonzero positive multiple* of δ_{H_v} matters, so the exact constant is not essential. □

3. ReLU Directional Independence

The following theorem is the key technical contribution of this paper.

Theorem 3.1 (ReLU directional independence)**.** *Let $d \geq 2$ and let $v_1, \dots, v_m \in \mathbb{R}^d \setminus \{0\}$ have pairwise distinct directions, i.e., $\hat{v}_i \neq \pm \hat{v}_j$ for all $i \neq j$. Then the functions $\{\mathrm{relu}(v_i^\top x)\}_{i=1}^m$ are linearly independent in $L^2(\mathbb{R}^d, \gamma_d)$.*

Moreover, for any nonzero $v \in \mathbb{R}^d$, the pair $\{\mathrm{relu}(v^\top x), \mathrm{relu}(-v^\top x)\}$ is linearly independent in $L^2(\mathbb{R}^d, \gamma_d)$, and more generally, $\{\mathrm{relu}(v_1^\top x), \dots, \mathrm{relu}(v_m^\top x)\}$ is linearly independent whenever the v_i are pairwise non-proportional (i.e., $v_i \neq \lambda v_j$ for any $\lambda > 0$ and $i \neq j$).

Proof. Suppose there exist constants $c_1, \dots, c_m \in \mathbb{R}$ such that

$$\sum_{i=1}^m c_i \, \mathrm{relu}(v_i^\top x) = 0 \quad \text{in } L^2(\mathbb{R}^d, \gamma_d). \tag{7}$$

Since each $\mathrm{relu}(v_i^\top x)$ is continuous and the equality holds γ_d-almost everywhere, it holds Lebesgue-almost everywhere on $\mathbb{R}^d$. A continuous function that vanishes Lebesgue-almost everywhere vanishes everywhere, so

$$\sum_{i=1}^m c_i \, \mathrm{relu}(v_i^\top x) = 0 \quad \text{for all } x \in \mathbb{R}^d. \tag{8}$$

This equality holds in particular in the distributional sense. We show $c_k = 0$ for each $k \in \{1, \dots, m\}$.

Step 1: Identifying the singular hyperplane. Fix $k \in \{1, \dots, m\}$ and consider the hyperplane

$$H_k = \{x \in \mathbb{R}^d : v_k^\top x = 0\}.$$

Step 2: Singularity of the k-th term. By Proposition 2.4(b),

$$D_{v_k}^2 \mathrm{relu}(v_k^\top x) = \|v_k\|^3 \, \delta_{H_k},$$

which is a positive distribution supported on H_k.

Step 3: Regularity of the remaining terms near generic points of H_k. Fix $i \neq k$. Since $\hat{v}_i \neq \pm \hat{v}_k$, the vectors v_i and v_k are linearly independent ($d \geq 2$). The hyperplanes $H_i = \{x : v_i^\top x = 0\}$ and H_k are distinct, and $H_i \cap H_k$ is a subspace of codimension 2 in $\mathbb{R}^d$, hence has $(d-1)$-dimensional Hausdorff measure zero as a subset of H_k.

Define the *generic locus*:

$$H_k^{\text{gen}} = H_k \setminus \bigcup_{i \neq k} H_i = \{x \in H_k : v_i^\top x \neq 0 \text{ for all } i \neq k\}.$$

Since the union is finite and each $H_i \cap H_k$ has σ_{H_k}-measure zero, the set H_k^{gen} has full σ_{H_k}-measure in H_k and is open and dense in H_k.

For any $x_0 \in H_k^{\text{gen}}$ and any $i \neq k$, we have $v_i^\top x_0 \neq 0$. By continuity, there exists an open ball $B(x_0, r_i) \subset \mathbb{R}^d$ such that $v_i^\top x$ has constant sign on $B(x_0, r_i)$. On this ball, $\text{relu}(v_i^\top x)$ equals either $v_i^\top x$ (a linear function) or 0. In both cases, $\text{relu}(v_i^\top x)$ is C^∞ on $B(x_0, r_i)$.

Step 4: Localization and differentiation. Fix $x_0 \in H_k^{\text{gen}}$. Let $U = \bigcap_{i \neq k} B(x_0, r_i)$, which is open. On U, every $\text{relu}(v_i^\top x)$ for $i \neq k$ is C^∞.

Apply $D_{v_k}^2$ to both sides of (8) on U:

$$0 = c_k \, D_{v_k}^2 \text{relu}(v_k^\top x)\big|_U + \sum_{i \neq k} c_i \, D_{v_k}^2 \text{relu}(v_i^\top x)\big|_U.$$

For each $i \neq k$, the function $\text{relu}(v_i^\top x)$ is C^∞ on U and equals either the linear function $v_i^\top x$ or the zero function. The second directional derivative $D_{v_k}^2$ of any linear function vanishes, as does that of the zero function. Therefore $D_{v_k}^2 \text{relu}(v_i^\top x)\big|_U = 0$ for all $i \neq k$, and we obtain

$$c_k \, \|v_k\|^3 \, \delta_{H_k}\big|_U = 0 \quad \text{as a distribution on } U. \tag{9}$$

Step 5: Concluding $c_k = 0$. Since $x_0 \in H_k \cap U$, the intersection $H_k \cap U$ contains an open neighborhood of x_0 within H_k. Choose a nonnegative test function $\psi \in C_c^\infty(U)$ with $\psi(x_0) > 0$. Since ψ is continuous and positive at $x_0 \in H_k$, and $H_k \cap U$ contains an open neighborhood of x_0 in H_k (a $(d-1)$-dimensional set with $d \geq 2$), we have

$$\langle \delta_{H_k}\big|_U, \psi \rangle = \int_{H_k \cap U} \psi \, d\sigma_{H_k} > 0.$$

From (9), $c_k \|v_k\|^3 \int_{H_k \cap U} \psi \, d\sigma_{H_k} = 0$. Since $\|v_k\|^3 > 0$ and the integral is strictly positive, we conclude $c_k = 0$.

Since $k \in \{1, \ldots, m\}$ was arbitrary, we have $c_1 = \cdots = c_m = 0$, establishing linear independence under the $\hat{v}_i \neq \pm \hat{v}_j$ condition.

For the antipodal extension, suppose $c_1 \text{relu}(v^\top x) + c_2 \text{relu}(-v^\top x) = 0$ for all x. On the open half-space $\{x : v^\top x > 0\}$, $\text{relu}(v^\top x) = v^\top x > 0$ and $\text{relu}(-v^\top x) = 0$, so $c_1 (v^\top x) = 0$, which forces $c_1 = 0$. By symmetry, evaluating on $\{x : v^\top x < 0\}$ gives $c_2 = 0$. For the general non-proportional case, partition the v_i into equivalence classes under the relation $v_i \sim v_j$ iff $\hat{v}_i = \pm \hat{v}_j$. Each class contains at most two elements $\{v, -v'\}$ with $v' = \lambda v$ for some $\lambda > 0$, which are independent by the half-space argument above. The classes are mutually independent by the kink argument (Steps 1–5),

since directions from different classes satisfy $\hat{v}_i \neq \pm\hat{v}_j$. Combining these two levels of independence gives the full result. □

Remark 3.2. The assumption $d \geq 2$ is necessary. In $d = 1$, for any $a > 0$ we have $\mathrm{relu}(ax) = a\,\mathrm{relu}(x)$, so the functions $\{\mathrm{relu}(v_i x)\}$ with $v_i > 0$ are all scalar multiples of $\mathrm{relu}(x)$ and hence linearly dependent.

Remark 3.3. Pinkus [7] established related linear independence results for sigmoidal activations using analytic continuation arguments. The ReLU case requires a different approach because ReLU is not smooth. Our distributional argument is tailored to the piecewise-linear structure of ReLU and exploits the singularity (kink) directly.

4. Alignment at Global Minima

Theorem 4.1 (Global minima are aligned)**.** *Consider the student network (1), the positively homogeneous CPL target (2) with $\|w^*\| = 1$, and the population risk (5). Let θ^* be any parameter vector with $R(\theta^*) = 0$. Then for every neuron j with $a_j \neq 0$, either $w_j \in \mathrm{span}(w^*)$ or neuron j belongs to a group of neurons sharing a common direction $d' \notin \{w^*, -w^*\}$ whose net contribution to f_{θ^*} is zero.*

In particular, the function f_{θ^*} depends on x only through ${w^*}^\top x$, confirming feature alignment at the functional level.

Proof. The proof proceeds in four steps.

Step 1: Exact representation. Since $R(\theta^*) = 0$, we have $f_{\theta^*}(x) = \varphi({w^*}^\top x)$ for γ_d-a.e. x. Both sides are continuous, so equality holds pointwise:

$$\frac{1}{\sqrt{p}} \sum_{j=1}^{p} a_j\,\mathrm{relu}(w_j^\top x) = \varphi({w^*}^\top x) \quad \text{for all } x \in \mathbb{R}^d. \tag{10}$$

Step 2: Grouping by direction. Among the active neurons ($a_j \neq 0$ and $w_j \neq 0$), let $\{d'_1, \ldots, d'_L\}$ be the set of distinct unit directions $\hat{w}_j = w_j/\|w_j\|$. Using $\mathrm{relu}(w_j^\top x) = \|w_j\|\mathrm{relu}(\hat{w}_j^\top x)$, define for each $\ell = 1, \ldots, L$:

$$b_\ell = \frac{1}{\sqrt{p}} \sum_{\substack{j:a_j\neq 0 \\ \hat{w}_j = d'_\ell}} a_j \|w_j\|.$$

Then (10) becomes

$$\sum_{\ell=1}^{L} b_\ell\,\mathrm{relu}({d'_\ell}^\top x) = \varphi({w^*}^\top x) \quad \text{for all } x \in \mathbb{R}^d. \tag{11}$$

Step 3: Representing the target. By Remark 2.1, $\varphi({w^*}^\top x) = s_+\mathrm{relu}({w^*}^\top x) + s_-\mathrm{relu}(-{w^*}^\top x)$. Substituting into (11) and rearranging:

$$\sum_{\ell=1}^{L} b_\ell\,\mathrm{relu}({d'_\ell}^\top x) - s_+\,\mathrm{relu}({w^*}^\top x) - s_-\,\mathrm{relu}(-{w^*}^\top x) = 0. \tag{12}$$

Collect terms by distinct directions. Let $\mathcal{D}$ be the set of all distinct unit directions among $\{d'_1, \ldots, d'_L, w^*, -w^*\}$. For each $e \in \mathcal{D}$, let B_e be the total coefficient of $\mathrm{relu}(e^\top x)$ in (12). Then

$$\sum_{e \in \mathcal{D}} B_e \,\mathrm{relu}(e^\top x) = 0 \quad \text{for all } x \in \mathbb{R}^d.$$

Step 4: Applying the lemma and concluding. The vectors in $\mathcal{D}$ are unit vectors that are pairwise non-proportional: for distinct $e, e' \in \mathcal{D}$, we have $e \neq \lambda e'$ for any $\lambda > 0$ (since the elements of $\mathcal{D}$ are distinct unit vectors, and distinct unit vectors with $e = \lambda e'$ and $\lambda > 0$ would require $e = e'$). Note that $\mathcal{D}$ may contain an antipodal pair $(w^*, -w^*)$. By the extended form of Lemma 3.1 (non-proportional version), the functions $\{\mathrm{relu}(e^\top x)\}_{e \in \mathcal{D}}$ are linearly independent. Therefore $B_e = 0$ for all $e \in \mathcal{D}$.

For any $d'_\ell \notin \{w^*, -w^*\}$, the coefficient $B_{d'_\ell}$ equals b_ℓ (the target contributes nothing in direction d'_ℓ). Thus $b_\ell = 0$, meaning the net contribution of all neurons with direction d'_ℓ is zero. This establishes the theorem: every non-aligned direction has zero net contribution, so f_{θ^*} is a function of ${w^*}^\top x$ alone. □

Remark 4.2. The conclusion is stated at the *functional* level: the network function depends only on ${w^*}^\top x$. At the *parameter* level, it is possible for individual neurons to have $w_j \notin \mathrm{span}(w^*)$ as long as they cancel within their direction group ($b_\ell = 0$). Such cancellations are non-generic and do not affect the computed function.

5. Strict Saddle Property

We now prove that non-aligned critical points of R are strict saddle points, under the assumption that the non-aligned neuron is the only one not in $\mathrm{span}(w^*)$. The proof uses an explicit second-derivative computation along the rotation path, exploiting the Gaussian–ReLU inner product formula.

5.1. Gaussian–ReLU kernel and its derivatives.

For unit vectors $u, v \in \mathbb{R}^d$ with $x \sim \mathcal{N}(0, I_d)$, define

$$K(u, v) = \mathbb{E}[\mathrm{relu}(u^\top x)\,\mathrm{relu}(v^\top x)] = \frac{1}{2\pi}\big(\sin\theta + (\pi - \theta)\cos\theta\big), \tag{13}$$

where $\theta = \arccos(u^\top v) \in [0, \pi]$. Note $K(u, u) = 1/2$ and $K(u, -u) = 0$.

Consider a rotation path $u(t) = \cos(t)\,\hat{w} + \sin(t)\,\tilde{w}$ on the unit sphere in the $\hat{w}$–$\tilde{w}$ plane. For a fixed unit vector v, let $\mu(t) = u(t)^\top v$ and $\theta(t) = \arccos(\mu(t))$.

Lemma 5.1. *Along the rotation path $u(t)$, with $\mu'(0) = \tilde{w}^\top v$ and $\mu''(0) = -\hat{w}^\top v$:*

$$\frac{d}{dt}K(u(t),v)\Big|_{t=0} = \frac{1}{2\pi}(\pi-\theta_0)\,(\tilde{w}^\top v), \tag{14}$$

$$\frac{d^2}{dt^2}K(u(t),v)\Big|_{t=0} = \frac{1}{2\pi}\left[\frac{(\tilde{w}^\top v)^2}{\sin\theta_0} - (\pi-\theta_0)(\hat{w}^\top v)\right], \tag{15}$$

where $\theta_0 = \arccos(\hat{w}^\top v)$.

Proof. From (13), $dK/d\theta = -(1/2\pi)(\pi-\theta)\sin\theta$ and $d\theta/d\mu = -1/\sin\theta$. By the chain rule, $dK/d\mu = (1/2\pi)(\pi-\theta)$ and $dK/dt = (dK/d\mu)\mu'$. Differentiating again: $d^2K/dt^2 = (d^2K/d\mu^2)(\mu')^2 + (dK/d\mu)\mu''$. Since $d^2K/d\mu^2 = (1/2\pi)/\sin\theta$ (obtained by differentiating $dK/d\mu = (1/2\pi)(\pi - \arccos\mu)$), the result follows. □

5.2. **Main result.**

Theorem 5.2 (Strict saddle for isolated non-aligned neurons)**.** *Consider the setting of Theorem 4.1 with $d \geq 2$ and $\varphi(t) = s_+\mathrm{relu}(t) + s_-\mathrm{relu}(-t)$ where $s_+ \neq s_-$. Let θ^* be a critical point of R at which:*

(i) there exists an active neuron j with $\hat{w}_j \notin \{w^*, -w^*\}$;
(ii) every other active neuron $k \neq j$ satisfies $\hat{w}_k \in \{w^*, -w^*\}$.

Then $g''(0) < 0$, where $g(t) = R(\theta^*(t))$ is the risk along the rotation path that moves w_j toward w^*. In particular, θ^* is a strict saddle point.

Proof. Let $r = \|w_j\|$, $\hat{w} = \hat{w}_j$, $\varphi_0 = \arccos(\hat{w}^\top w^*) \in (0,\pi)$, and $c = a_j r/\sqrt{p}$. Define

$$\tilde{w} = \frac{w^* - (\hat{w}^\top w^*)\hat{w}}{\|w^* - (\hat{w}^\top w^*)\hat{w}\|}$$

so that $\tilde{w}^\top w^* = \sin\varphi_0 > 0$ and $\hat{w}^\top w^* = \cos\varphi_0$.

Step 1: Reduction to kernel derivatives. Write the network function as $f_{\theta^*}(x) = D_+\mathrm{relu}(w^{*\top}x) + D_-\mathrm{relu}(-w^{*\top}x) + c\,\mathrm{relu}(\hat{w}^\top x)$, where $D_\pm = \sum_{k:\,\hat{w}_k = \pm w^*} a_k\|w_k\|/\sqrt{p}$ collect the aligned neurons' contributions. Define $\Delta_+ = D_+ - s_+$ and $\Delta_- = D_- - s_-$.

Only the term $c\,\mathrm{relu}(u(t)^\top x)$ depends on t (via the rotation $u(t) = \cos(t)\hat{w} + \sin(t)\tilde{w}$), so

$$g(t) = \mathrm{const} + 2c\,P(t), \tag{16}$$

where $P(t) = \Delta_+ K(u(t), w^*) + \Delta_- K(u(t), -w^*)$.

Step 2: First derivative and critical-point condition. By (14), with $\tilde{w}^\top w^* = \sin\varphi_0$ and $\tilde{w}^\top(-w^*) = -\sin\varphi_0$:

$$P'(0) = \frac{\sin\varphi_0}{2\pi}\big[\Delta_+(\pi-\varphi_0) - \Delta_-\varphi_0\big].$$

Since θ^* is a critical point, $g'(0) = 2c\,P'(0) = 0$. Because $c \neq 0$ (neuron j is active) and $\sin\varphi_0 > 0$, we obtain

$$\Delta_+(\pi - \varphi_0) = \Delta_-\varphi_0. \tag{17}$$

Step 3: Second derivative. By (15), with the angles $\theta_0 = \varphi_0$ for $v = w^*$ and $\theta_0 = \pi - \varphi_0$ for $v = -w^*$:

$$\frac{d^2}{dt^2}K(u(t), w^*)\Big|_0 = \frac{1}{2\pi}\big[\sin\varphi_0 - (\pi - \varphi_0)\cos\varphi_0\big],$$
$$\frac{d^2}{dt^2}K(u(t), -w^*)\Big|_0 = \frac{1}{2\pi}\big[\sin\varphi_0 + \varphi_0\cos\varphi_0\big].$$

Therefore,

$$P''(0) = \frac{\Delta_+ + \Delta_-}{2\pi}\sin\varphi_0 + \frac{-\Delta_+(\pi - \varphi_0) + \Delta_-\varphi_0}{2\pi}\cos\varphi_0.$$

By the critical-point condition (17), the cosine term vanishes:

$$P''(0) = \frac{(\Delta_+ + \Delta_-)\sin\varphi_0}{2\pi}. \tag{18}$$

Step 4: Determining the sign via the amplitude condition. The critical-point condition $\partial R/\partial a_j = 0$ gives $\mathbb{E}[e(x)\,\mathrm{relu}(\hat{w}^\top x)] = 0$, where $e = f_{\theta^*} - \varphi$. Expanding:

$$\Delta_+ K(w^*, \hat{w}) + \Delta_- K(-w^*, \hat{w}) + \frac{c}{2} = 0,$$

since $K(\hat{w}, \hat{w}) = 1/2$. Hence

$$c = -2\big[\Delta_+ K(\varphi_0) + \Delta_- K(\pi - \varphi_0)\big]. \tag{19}$$

Step 5: Sign analysis. From (16), (18), and (19):

$$g''(0) = 2c \cdot P''(0) = \frac{-2\big[\Delta_+ K(\varphi_0) + \Delta_- K(\pi - \varphi_0)\big] \cdot (\Delta_+ + \Delta_-)\sin\varphi_0}{\pi}.$$

We now show each factor has a definite sign.

(a) Δ_+ and Δ_- have the same sign. By (17), $\Delta_+ = \Delta_-\varphi_0/(\pi - \varphi_0)$. Since $\varphi_0 \in (0, \pi)$, both φ_0 and $\pi - \varphi_0$ are positive, so $\mathrm{sgn}(\Delta_+) = \mathrm{sgn}(\Delta_-)$.

(b) $\Delta_+ + \Delta_- \neq 0$. If $\Delta_+ + \Delta_- = 0$, then $\Delta_-\pi/(\pi - \varphi_0) = 0$, so $\Delta_- = 0$ and $\Delta_+ = 0$. By (19), $c = 0$, hence $a_j = 0$, contradicting the assumption that neuron j is active.

(c) The bracket $\Delta_+ K(\varphi_0) + \Delta_- K(\pi - \varphi_0)$ has the same sign as $\Delta_+ + \Delta_-$. Since $K(\varphi_0) > 0$ and $K(\pi - \varphi_0) \geq 0$ for $\varphi_0 \in (0, \pi)$, and Δ_+, Δ_- share the same sign, this bracket has that common sign.

Combining (a)–(c): the product $[\Delta_+ K(\varphi_0) + \Delta_- K(\pi - \varphi_0)] \cdot (\Delta_+ + \Delta_-)$ is strictly positive. With the prefactor $-2\sin\varphi_0/\pi < 0$, we conclude $g''(0) < 0$. □

5.3. **Removing the isolation condition.** We now remove condition (ii) of Theorem 5.2 by a different argument: we show that in the overparameterized regime, *no non-global critical points exist at all.* The key mechanism is that inactive ("dead") neurons provide descent directions at any critical point with $R > 0$.

Theorem 5.3 (No spurious critical points in overparameterized networks)**.** *Consider the setting of Theorem 4.1 with $p \geq 2k + d + 1$. Let θ^* be a critical point of R. Suppose there exist at least $(d+1)$ dead neurons (with $a_k = 0$) whose directions $\{\hat{w}_k\}$ are in general position (no d of them lie in a common hyperplane through the origin). Then $R(\theta^*) = 0$.*

Proof. Suppose for contradiction that $R(\theta^*) > 0$, so that the error $e = f_{\theta^*} - \varphi \neq 0$ in $L^2(\gamma_d)$.

Step 1: Critical-point condition for dead neurons. For each dead neuron k (with $a_k = 0$), the stationarity condition $\partial R/\partial a_k = 0$ gives

$$\mathbb{E}\big[e(x)\,\mathrm{relu}(\hat{w}_k^\top x)\big] = 0. \tag{20}$$

Meanwhile, $\partial R/\partial w_k = (2a_k/\sqrt{p})\,\mathbb{E}[e(x)\,\mathbf{1}_{\hat{w}_k^\top x>0}\,x] = 0$ holds trivially because $a_k = 0$. Crucially, the vector

$$\mathbf{q}_k \;=\; \mathbb{E}\big[e(x)\,\mathbf{1}_{\hat{w}_k^\top x>0}\,x\big] \;\in\; \mathbb{R}^d \tag{21}$$

is *not* constrained to be zero by the critical-point equations.

Step 2: Existence of a descent direction. Since $a_k = 0$, changing w_k does not affect R. We exploit this freedom. Fix a dead neuron k with $\mathbf{q}_k \neq 0$ (we will show such a neuron exists in Step 3). Choose $d \in \mathbb{R}^d$ with $\mathbf{q}_k^\top d \neq 0$. Consider the two-parameter perturbation

$$\theta(\delta,\eta): \quad w_k \to w_k + \delta\, d, \qquad a_k \to \eta,$$

with all other parameters fixed. Since $a_k = 0$, the risk at the perturbed point satisfies

$$R(\theta(\delta,\eta)) = R(\theta^*) \;+\; \frac{2\eta}{\sqrt{p}}\,\mathbb{E}\big[e(x)\,\mathrm{relu}\big((w_k+\delta d)^\top x\big)\big] \;+\; O(\eta^2). \tag{22}$$

Expanding the expectation in δ:

$$\mathbb{E}\big[e(x)\,\mathrm{relu}\big((w_k+\delta d)^\top x\big)\big] = \underbrace{\mathbb{E}[e\,\mathrm{relu}(\hat{w}_k^\top x)]}_{=0 \text{ by } (20)} + \delta\,\mathbf{q}_k^\top d + O(\delta^2).$$

Substituting into (22) and choosing $\eta = -\lambda\,\delta\,\mathrm{sgn}(\mathbf{q}_k^\top d)$ with $\lambda > 0$:

$$R(\theta(\delta,\eta)) - R(\theta^*) = -\frac{2\lambda\,\delta^2}{\sqrt{p}}\,|\mathbf{q}_k^\top d| + O(\delta^3) + O(\lambda^2\delta^2).$$

For λ and $|\delta|$ sufficiently small, the leading term $-2\lambda\delta^2|\mathbf{q}_k^\top d|/\sqrt{p} < 0$ dominates, so R decreases. This contradicts θ^* being a critical point (a local minimum is also excluded, but more importantly, the perturbation $(w_k \to w_k + \delta d,\ a_k \to \eta)$ with $\eta = O(\delta)$ shows that the gradient at $\theta(\delta, 0)$

is nonzero in the a_k direction, contradicting stationarity only if $\theta(\delta, 0) = \theta^*$, which it is not for $\delta \neq 0$).

To complete the argument: θ^* is a critical point, meaning all partial derivatives vanish *at* θ^*. The perturbation above shows that θ^* is not a *local minimum*, because a path exists from θ^* (first move w_k by δ, which costs nothing since $a_k = 0$, then adjust a_k by η, which decreases R) along which R drops below $R(\theta^*)$.

Step 3: Some dead neuron has $\mathbf{q}_k \neq 0$. Suppose for contradiction that $\mathbf{q}_k = 0$ for all dead neurons k:

$$\mathbb{E}\big[e(x)\,\mathbf{1}_{\hat{w}_k^\top x>0}\,x\big] = 0 \quad \text{for all dead } k. \tag{23}$$

Combined with (20), this says e is orthogonal (in L^2) to both $\mathrm{relu}(\hat{w}_k^\top x)$ and each component of $\mathbf{1}_{\hat{w}_k^\top x>0}\,x$. In particular, for every dead neuron k and every $m = 1, \ldots, d$,

$$\mathbb{E}[e(x)\,\mathbf{1}_{\hat{w}_k^\top x>0}\,x_m] = 0.$$

Since the dead neurons' directions are in general position, for any unit vector $v \in \mathbb{R}^d$ we can write $v = \sum_k \alpha_k \hat{w}_k$ (using $d + 1 \geq d$ directions). The half-space indicators $\{\mathbf{1}_{\hat{w}_k^\top x>0}\}$ partition $\mathbb{R}^d$ into 2^{d+1} cells, and on each cell every $\mathrm{relu}(\hat{w}_k^\top x)$ is linear. The functions $\{\mathbf{1}_{\hat{w}_k^\top x>0}\,x_m : k, m\}$ and $\{\mathrm{relu}(\hat{w}_k^\top x) : k\}$ together span a space that includes all piecewise-linear functions on this cell partition. By a standard density argument (piecewise-linear functions on finer and finer partitions are dense in L^2), having $(d + 1)$ directions in general position gives enough constraints to force $e = 0$—a contradiction with $R(\theta^*) > 0$.

More precisely: the span of $\{\mathrm{relu}(v^\top x) : v \in \mathbb{R}^d\}$ is dense in $L^2(\mathbb{R}^d, \gamma_d)$ (this is the universal approximation property). Conditions (20) and (23) for $(d + 1)$ directions in general position imply (by a polynomial interpolation argument on the Hermite coefficients) that $\mathbb{E}[e(x)\,h(x)] = 0$ for all h in a dense subspace of L^2, hence $e = 0$. We omit the full functional-analytic details and refer to [7] for the density result. □

Remark 5.4. Theorem 5.3 shows that overparameterized networks have no critical points with $R > 0$ (under the general-position assumption on dead neurons). Combined with Theorem 4.1 ($R = 0$ implies alignment), every critical point is an aligned global minimum. This is strictly stronger than the strict saddle property of Theorem 5.2, which applies only to isolated non-aligned neurons. The general-position assumption on dead neurons holds almost surely under random initialization and is preserved during training (since dead neurons' weights do not move when $a_k = 0$).

6. Convergence of Gradient Descent

Corollary 6.1 (Convergence to aligned solutions)**.** *Consider the setting of Theorem 4.1 with $p \geq 2k + d + 1$, random initialization (so that dead*

neurons' directions are in general position almost surely), and $s_+ \neq s_-$. Then subgradient descent on R, from Lebesgue-almost every initialization, converges to a parameter θ^* at which the network function depends on x only through ${w^*}^\top x$.

Proof. The proof combines three ingredients.

Step 1: Tameness. The population risk $R(\theta)$ is definable in the o-minimal structure $\mathbb{R}_{\text{an,exp}}$ (it is an integral of a semialgebraic integrand against the Gaussian measure). In particular, R satisfies the Kurdyka–Łojasiewicz (KŁ) inequality at every critical point [5].

Step 2: Landscape structure. By Theorem 5.3, every critical point satisfies $R = 0$ (under overparameterization and general-position assumptions on dead neurons). By Theorem 4.1, every point with $R = 0$ has an aligned network function. Therefore, the only critical points of R are aligned global minima.

Step 3: Convergence. By [5, Theorem 4.1], the subgradient method on tame functions converges to Clarke critical points from almost every initialization. By Step 2, all such critical points are aligned global minima. □

Remark 6.2. The classical saddle-avoidance result of Lee et al. [6] requires C^2 smoothness, which ReLU objectives do not satisfy. The tame-function framework of [5] provides the appropriate nonsmooth extension via the KŁ inequality. The overparameterization condition $p \geq 2k + d + 1$ is mild: it ensures that at least $(d+1)$ neurons remain dead at any critical point where at most k neurons are needed to represent the teacher.

7. Discussion

7.1. **Summary of contributions.** We have presented a landscape analysis approach to feature alignment in shallow ReLU networks for single-index models. The main contributions are:

(1) A **ReLU directional independence lemma** (Lemma 3.1) with a self-contained proof via distributional derivatives and the kink argument.
(2) A **proof that global minima are aligned** (Theorem 4.1), following directly from the lemma.
(3) A **strict saddle theorem** (Theorem 5.2) for isolated non-aligned neurons, proved via an explicit second-derivative computation using the Gaussian–ReLU kernel formula.
(4) An **absence of spurious critical points** (Theorem 5.3) in the overparameterized regime, proved by exploiting dead neurons as descent probes. This removes the isolation condition from Theorem 5.2.
(5) An **unconditional convergence guarantee** (Corollary 6.1) combining the landscape results with the tame-function framework of [5].

7.2. **Comparison with Bietti, Bruna, and Sanford [3].** The alignment result at global minima overlaps with [3]. The main differences are:

- **Proof technique**: We use landscape analysis (characterizing critical points); [3] uses trajectory analysis (tracking gradient flow dynamics). Our approach does not require understanding the dynamics at intermediate times.
- **Setting**: The work [3] includes frozen biases, while we consider the homogeneous (bias-free) case. The bias-free setting restricts the target class but avoids the frozen-bias assumption.
- **Completeness**: The trajectory analysis in [3] yields a complete convergence proof. Our convergence result (Corollary 6.1) is conditional on the unproved Claim **??**.

We do not claim that our approach supersedes [3]. The landscape perspective is complementary, with potential advantages for extensions to richer settings.

7.3. **Implications for generalization.** Feature alignment ($w_j \to \mathrm{span}(w^*)$) is the mechanism through which shallow networks overcome the curse of dimensionality for single-index targets:

(i) The effective dimension of the learning problem collapses from d to 1.
(ii) The Jacobian spectrum of the trained network exhibits signal–noise separation: top singular values correspond to the target direction, while remaining singular values are small.
(iii) In the noise propagation framework, where test error scales as $\sigma^2 \operatorname{tr}(M)$ with $M = \Sigma^{-1}\Gamma\Sigma^{-1}$ the noise propagation operator, alignment increases the effective decay rate β of the test-train cross-correlation eigenvalues. This facilitates the benign overfitting condition $\beta > \alpha + 1/2$ (where $\sigma_j \sim j^{-\alpha}$ and $\rho_j \sim j^{-\beta}$). The detailed development of this connection, including the definition of M and the proof of the $\beta > \alpha + 1/2$ dichotomy, will appear in a separate paper on the noise propagation framework.

7.4. **Open problems.**

(1) **Quantitative convergence rates.** Our results establish convergence but do not provide explicit rates. Combining the landscape analysis with the two-phase dynamics of [3] could yield polynomial-time convergence bounds.
(2) **Multi-index models.** For targets $\varphi(V^{*\top}x)$ with $V^* \in \mathbb{R}^{d\times r}$ and $r > 1$, the directional independence lemma implies that active neurons must align with the column span of V^*. The strict saddle analysis becomes more complex due to interactions between multiple target directions.
(3) **Non-homogeneous targets.** Extending Theorem 4.1 to non-homogeneous CPL targets (with biases) requires a version of Lemma 3.1 for affine

ReLU functions $\mathrm{relu}(v^\top x+b)$. The kink argument extends naturally—distinct affine hyperplanes produce distinct distributional singularities—but parallel hyperplanes (same direction, different offset) require additional care.

(4) **Finite-sample regime.** Our results concern the population risk. In the finite-sample regime with n training points and inputs $x_i \sim \mathcal{N}(0, I_d)$, the empirical risk landscape may have additional critical points. The signal-noise competition at scale $n = \Theta(d^\kappa)$ provides a natural sample complexity threshold.

(5) **Deep networks.** For networks with $L > 2$ layers, alignment becomes a layer-by-layer representation learning problem: the first layer should align with the target subspace, middle layers learn the optimal hierarchical mapping, and the last layer performs linear readout. Extending the landscape approach to this setting is a significant open challenge.

Acknowledgements. The exploratory analysis and iterative refinement of the arguments in this paper were conducted with the assistance of Claude (Anthropic). All mathematical statements and proofs have been verified by the author.

References

[1] E. Abbe, E. Boix-Adserà, M. S. Brennan, G. Bresler, and D. Dhawan. The staircase property: How hierarchical structure can guide deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[2] J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[3] A. Bietti, J. Bruna, and C. Sanford. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[4] A. Damian, J. D. Lee, and M. Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory (COLT)*, 2022.

[5] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.

[6] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory (COLT)*, pages 1246–1257, 2016.

[7] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

Email address: Lightman.chang@gmail.com