Globally Optimal Gradient Descent for a ConvNet with Gaussian Inputs

Supplementary Material

Contents

A Proof of Lemma 3.2

First assume that $\theta_{\boldsymbol{u},\boldsymbol{v}} \neq 0, \pi$. Then we have

$$\begin{gathered}
\frac{\partial g}{\partial u_i} = \frac{1}{2\pi}\left\|\boldsymbol{v}\right\|\frac{u_i}{\left\|\boldsymbol{u}\right\|}\left(\sqrt{1-\Big(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\Big)^2}+\Big(\pi-\arccos\Big(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\Big)\Big)\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\right)+ \\
\frac{1}{2\pi}\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|\left(\left(\quad-\frac{\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}}{\sqrt{1-\Big(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\Big)^2}}\right)\left(\frac{v_i}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}-\frac{u_i}{\left\|\boldsymbol{u}\right\|^2}\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\right)+\right. \\
\left(\frac{\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}}{\sqrt{1-\big(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\big)^2}}\left(\frac{v_i}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}-\frac{u_i}{\left\|\boldsymbol{u}\right\|^2}\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\right)\right)+ \\
\left.\Big(\pi-\arccos\Big(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\Big)\Big)\left(\frac{v_i}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}-\frac{u_i}{\left\|\boldsymbol{u}\right\|^2}\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\right)\right)= \\
\frac{1}{2\pi}\left\|\boldsymbol{v}\right\|\frac{u_i}{\left\|\boldsymbol{u}\right\|}\left(\sqrt{1-\Big(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\Big)^2}+\Big(\pi-\arccos\Big(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\Big)\Big)\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\right)+ \\
\frac{1}{2\pi}\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|\Big(\pi-\arccos\Big(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\Big)\Big)\left(\frac{v_i}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}-\frac{u_i}{\left\|\boldsymbol{u}\right\|^2}\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\right)\Bigg)=
\end{gathered}$$

$$\frac{1}{2\pi}\left\|\boldsymbol{v}\right\|\frac{u_i}{\left\|\boldsymbol{u}\right\|}\sqrt{1-\Big(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\Big)^2}+\frac{1}{2\pi}\Big(\pi-\arccos\Big(\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\left\|\boldsymbol{u}\right\|\left\|\boldsymbol{v}\right\|}\Big)\Big)v_i=$$

$$\frac{1}{2\pi}\left\|\boldsymbol{v}\right\|\frac{u_i}{\left\|\boldsymbol{u}\right\|}\sin\theta_{\boldsymbol{u},\boldsymbol{v}}+\frac{1}{2\pi}\Big(\pi-\theta_{\boldsymbol{u},\boldsymbol{v}}\Big)v_i$$

Hence,

$$\frac{\partial g}{\partial \boldsymbol{u}}=\frac{1}{2\pi}\left\|\boldsymbol{v}\right\|\frac{\boldsymbol{u}}{\left\|\boldsymbol{u}\right\|}\sin\theta_{\boldsymbol{u},\boldsymbol{v}}+\frac{1}{2\pi}\Big(\pi-\theta_{\boldsymbol{u},\boldsymbol{v}}\Big)\boldsymbol{v} \tag{1}$$

Now we assume that $\boldsymbol{u}$ is parallel to $\boldsymbol{v}$. We first show that g is differentiable in this case. Without loss of generality we can assume that $\boldsymbol{u}$ and $\boldsymbol{v}$ lie on the u_1 axis. This follows since g is a function of $\left\|\boldsymbol{u}\right\|$, $\left\|\boldsymbol{v}\right\|$ and $\theta_{\boldsymbol{u},\boldsymbol{v}}$ and therefore $g(\cdot,\boldsymbol{v})$ has a directional derivative in direction $\mathbf{d}$ at $\boldsymbol{u}$ if and only if $g(\cdot,R\boldsymbol{v})$ has a directional derivative in direction $R\mathbf{d}$ at $R\boldsymbol{u}$ where R is a rotation matrix. Hence $g(\cdot,\boldsymbol{v})$ is differentiable at $\boldsymbol{u}$ if and only if $g(\cdot,R\boldsymbol{v})$ is differentiable at $R\boldsymbol{u}$. Furthermore, if $\boldsymbol{v}$ and $\boldsymbol{u}$ are on the u_1 axis, then by symmetry the partial derivatives with respect to *other* axes at $\boldsymbol{u}$ are all equal, hence we only need to consider the partial derivative with respect to the u_1 and u_2 axes.

Let $\boldsymbol{v}=(1,0,...,0)$ and $\boldsymbol{u}=(u,0,...,0)$ where $u\neq 0$. In order to show differentiability, we will prove that $g(\boldsymbol{u},\boldsymbol{v})$ has continuous partial derivatives at $\boldsymbol{u}$ (by equality (1) the partial derivatives are clearly continuous at points that are not on the u_1 axis. Define $\boldsymbol{u}_\epsilon=(u,\epsilon,0,...,0)$. Then

$$\frac{\partial g}{\partial u_2}(\boldsymbol{u},\boldsymbol{v})=\lim_{\epsilon\to 0}\frac{\frac{1}{2\pi}\left\|\boldsymbol{u}_\epsilon\right\|\left\|\boldsymbol{v}\right\|\left(\sin\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{v}}+\Big(\pi-\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{v}}\Big)\cos\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{v}}\right)-g(\boldsymbol{u},\boldsymbol{v})}{\epsilon}$$

By L'hopital's rule and the calculation of equality (1) we get

$$\frac{\partial g}{\partial u_2}(\boldsymbol{u},\boldsymbol{v})=\lim_{\epsilon\to 0}\frac{1}{2\pi}\left\|\boldsymbol{v}\right\|\frac{\epsilon}{\left\|\boldsymbol{u}_\epsilon\right\|}\sin\theta_\epsilon=0$$

Furthermore, by equality (1) we see that $\lim_{\boldsymbol{u}'\to\boldsymbol{u}}\frac{\partial g}{\partial u_2}(\boldsymbol{u}',\boldsymbol{v})=0$ since $\lim_{\boldsymbol{u}'\to\boldsymbol{u}}\sin\theta_{\boldsymbol{u}',\boldsymbol{v}}=0$.

For a fixed $\theta_{\boldsymbol{u},\boldsymbol{v}}$ equal to 0 or π, $\frac{\partial g}{\partial u_1}(\boldsymbol{u},\boldsymbol{v})$ is the same as $\frac{\partial g}{\partial\left\|\boldsymbol{u}\right\|}(\boldsymbol{u},\boldsymbol{v})$. Hence,

$$\frac{\partial g}{\partial u_1}(\boldsymbol{u},\boldsymbol{v})=\frac{1}{2\pi}\left\|\boldsymbol{v}\right\|\left(\sin\theta_{\boldsymbol{u},\boldsymbol{v}}+\Big(\pi-\theta_{\boldsymbol{u},\boldsymbol{v}}\Big)\cos\theta_{\boldsymbol{u},\boldsymbol{v}}\right)=\begin{cases}\frac{1}{2} \text{ if } u>0\\ 0 \text{ if } u<0\end{cases}$$

and the partial derivative is continuous since

$$\lim_{\boldsymbol{u}'\to\boldsymbol{u}}\frac{\partial g}{\partial u_1}(\boldsymbol{u}',\boldsymbol{v})=\begin{cases}\frac{1}{2} \text{ if } u>0\\ 0 \text{ if } u<0\end{cases}$$

Finally, we see that for the case where $\boldsymbol{u}$ and $\boldsymbol{v}$ are parallel, the values we got for the partial derivatives coincide with equation Eq. 1. This concludes the proof.

B Proof of Proposition 4.1

We will prove the claim by induction on k. For the base case we will show that *Set-Splitting-by-2-Sets* is NP-complete. We will prove this via a reduction from a variant of the 3-SAT problem with the restriction of equal number of variables and clauses, which we denote *Equal-3SAT*. We will first prove that *Equal-3SAT* is NP-complete.

Lemma B.1. *Equal-3SAT is NP-complete.*

Proof. This can be shown via a reduction from 3SAT. Given a formula ϕ with n variables and m clauses we can increase $n-m$ by 1 by adding a new clause of the form $(x \vee y)$ for new variables x and y. Furthermore, we can decrease $n-m$ by 1 by adding two new identical clauses of the form (z) for a new variable z. In each case the formula with the new clause(s) is satisfiable if and only if ϕ is. Therefore given a formula ϕ we can construct a new formula ψ with equal number of variables and clauses such that ϕ is satisfiable if and only if ψ is. □

We will now give a reduction from *Equal-3SAT* to *Set-Splitting-by-2-Sets*.

Lemma B.2. *Set-Splitting-by-2-Sets is NP-complete.*

Proof. The following reduction is exactly the reduction from 3SAT to Splitting-Sets and we include it here for completeness. Let ϕ be a formula with set of variables V and equal number of variables and clauses. We construct the sets S and $\mathcal{C}$ as follows. Define

$$S = \{\bar{x} \mid x \in V\} \cup V \cup \{n\}$$

where $\bar{x}$ is the negation of variable x and n is a new variable not in V. For each clause c with set of variables or negations of variables V_c that appear in the clause (for example, if $c = (\bar{x} \vee y)$ then $V_c = \{\bar{x}, y\}$) construct a set $S_c = V_c \cup \{n\}$. Furthermore, for each variable $x \in V$ construct a set $S_x = \{x, \bar{x}\}$. Let $\mathcal{C}$ be the family of subsets S_c and S_x for all clauses c and $x \in V$. Note that $|\mathcal{C}| \leq |S|$ which is required by the definition of *Set-Splitting-by-2-Sets*.

Assume that ϕ is satisfiable and let A be the satisfying assignment. Define $S_1 = \{x | A(x) = true\} \cup \{\bar{x} | A(x) = false\}$ and $S_2 = \{x | A(x) = false\} \cup \{\bar{x} | A(x) = true\} \cup \{n\}$. Note that $S_1 \cup S_2 = S$. Assume by contradiction that there exists a set $T \in \mathcal{C}$ such that $T \subseteq S_1$ or $T \subseteq S_2$. If $T \subseteq S_1$ then T is not a set S_c for some clause c because $n \notin S_1$. However, by the construction of S_1 a variable and its negation cannot be in S_1. Hence $T \subseteq S_1$ is impossible. If $T \subseteq S_2$ then as in the previous claim T cannot be a set S_x for a variable x. Hence $T = S_c$ for some clause c. However, this implies that $A(c) = false$, a contradiction.

Conversely, assume there exists splitting sets S_1 and S_2 and w.l.o.g. $n \in S_1$. We note that it follows that no variable x and its negation $\bar{x}$ are both contained in one of the sets S_1 or S_2. Define the following assignment A for ϕ. For all $x \in V$ if $x \in S_1$ let $A(x) = false$, otherwise let $A(x) = true$. Note that A is a well defined assignment. Assume by contradiction that there is a clause c in ϕ which is not satisfiable. Since S_2 splits S_c it follows that there exists a variable x such that it or its negation $\bar{x}$ are in S_2 (recall that $n \in S_1$). If $x \in S_2$ then $A(x) = true$ and if $\bar{x} \in S_2$ then $A(\bar{x}) = true$ since $x \in S_1$. In both cases c is satisfiable, a contradiction. □

This proves the base case. We will now prove the induction step by giving a reduction from *Set-Splitting-by-k-Sets* to *Set-Splitting-by-(k+1)-Sets*. Given $S = \{1, 2, ..., d\}$ and $\mathcal{C} = \{C_j\}_j$ such that $|\mathcal{C}| \leq (k-1)d$, define $S' = \{1, 2, ..., d+1\}$ and $\mathcal{C}' = \mathcal{C} \cup \{D_j\}_j$ where $D_j = \{j, d+1\}$ for all $1 \leq j \leq d$. Note that $|\mathcal{C}'| \leq kd < k(d+1)$. Assume that there are $S_1, ..., S_k$ that split the sets in $\mathcal{C}$. Then if we define $S_{k+1} = \{d+1\}$, it follows that $\bigcup_{i=1}^{k+1} S_i = S$ and $S_1, ..., S_k, S_{k+1}$ are disjoint and split the sets in $\mathcal{C}'$.

Conversely, assume that $S_1, ..., S_k, S_{k+1}$ split the sets in $\mathcal{C}'$. Let w.l.o.g. S_{k+1} be the set that contains $d+1$. Then for all $1 \leq j \leq d$ we have $D_j \not\subseteq S_{k+1}$. It follows that for all $1 \leq j \leq d$, $j \notin S_{k+1}$, or equivalently, $S_{k+1} = \{d+1\}$. Hence, $\bigcup_{i=1}^{k} S_i = S$ and $S_1, ..., S_k$ are disjoint and split the sets in $\mathcal{C}$, as desired.

C Missing Proofs for Section 5

C.1 Proof of Proposition 5.1

1. For $\boldsymbol{w} \neq \boldsymbol{0}$, the claim follows from Lemma 3.2. As in the proof of Lemma 3.2 we can assume w.l.o.g. that $\boldsymbol{w} = (0, 0, ..., 0)$ and $\boldsymbol{w}^* = (1, 0, ..., 0)$. Let $f(\boldsymbol{w}, \boldsymbol{w}^*) = 2kg(\boldsymbol{w}, \boldsymbol{w}^*) + (k^2 -$

$k)\frac{\|\boldsymbol{w}\|\|\boldsymbol{w}^*\|}{\pi}$. It suffices to show that $\frac{\partial f}{\partial w_2}(\boldsymbol{w}, \boldsymbol{w}^*)$ does not exist. Indeed, let $\boldsymbol{w}_\epsilon = (0, \epsilon, 0, ..., 0)$ then by L'hopital's rule

$$\lim_{\epsilon \to 0^+} \frac{f(\boldsymbol{w}_\epsilon, \boldsymbol{w}^*) - f(\boldsymbol{w}, \boldsymbol{w}^*)}{\epsilon} = \lim_{\epsilon \to 0^+} \frac{k}{\pi} \|\boldsymbol{w}^*\| \frac{\epsilon}{|\epsilon|} \sin \theta_{\boldsymbol{w}_\epsilon, \boldsymbol{w}^*} + (k^2 - k) \frac{\|\boldsymbol{w}^*\|}{\pi} = \frac{k}{\pi} + \frac{k^2 - k}{\pi}$$

and

$$\lim_{\epsilon \to 0^-} \frac{f(\boldsymbol{w}_\epsilon, \boldsymbol{w}^*) - f(\boldsymbol{w}, \boldsymbol{w}^*)}{\epsilon} = \lim_{\epsilon \to 0^-} \frac{k}{\pi} \|\boldsymbol{w}^*\| \frac{\epsilon}{|\epsilon|} \sin \theta_{\boldsymbol{w}_\epsilon, \boldsymbol{w}^*} - (k^2 - k) \frac{\|\boldsymbol{w}^*\|}{\pi} = -\frac{k}{\pi} - \frac{k^2 - k}{\pi}$$

Hence the left and right partial derivatives with respect to variable w_2 are not equal, and thus $\frac{\partial f}{\partial w_2}(\boldsymbol{w}, \boldsymbol{w}^*)$ does not exist.

2. We first show that $\boldsymbol{w} = \mathbf{0}$ is a local maximum if and only if $k > 1$. Indeed, by considering the loss function as a function of the variable $x = \|\boldsymbol{w}\|$, for any *fixed* angle $\theta_{\boldsymbol{w}, \boldsymbol{w}^*}$ we get a quadratic function of the form $\ell(x) = ax^2 - bx$, where $a > 0$ and $b \geq 0$. Since $f(\theta) = \sin \theta + (\pi - \theta) \cos \theta$ is a non-negative function for $0 \leq \theta \leq \pi$ and $f(\theta) = 0$ if and only if $\theta = \pi$, it follows that $b = 0$ if and only if $k = 1$ *and* $\theta_{\boldsymbol{w}, \boldsymbol{w}^*} = \pi$. Therefore if $k > 1$, then for all fixed angles $\theta_{\boldsymbol{w}, \boldsymbol{w}^*}$, the minimum of $\ell(x)$ is attained at $x > 0$, which implies that $\boldsymbol{w} = \mathbf{0}$ is a local maximum. If $k = 1$ and $\theta_{\boldsymbol{w}, \boldsymbol{w}^*} = \pi$ the minimum of $\ell(x)$ is attained at $x = 0$, and thus $\boldsymbol{w} = \mathbf{0}$ is not a local maximum in this case.

   We will now find the other critical points of ℓ. By Lemma 3.2 we get

$$\begin{aligned} \nabla \ell(\boldsymbol{w}) &= \frac{1}{k^2} \left[(k + \frac{k^2 - k}{\pi}) \boldsymbol{w} - \frac{k}{\pi} \|\boldsymbol{w}^*\| \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \sin \theta_{\boldsymbol{w}, \boldsymbol{w}^*} - \frac{k}{\pi} \Big(\pi - \theta_{\boldsymbol{w}, \boldsymbol{w}^*} \Big) \boldsymbol{w}^* - \frac{k^2 - k}{\pi} \|\boldsymbol{w}^*\| \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \right] \\ &= \frac{1}{k^2} \left[\left(k + \frac{k^2 - k}{\pi} - \frac{k \|\boldsymbol{w}^*\|}{\pi \|\boldsymbol{w}\|} \sin \theta_{\boldsymbol{w}, \boldsymbol{w}^*} - \frac{k^2 - k}{\pi} \frac{\|\boldsymbol{w}^*\|}{\|\boldsymbol{w}\|} \right) \boldsymbol{w} - \frac{k}{\pi} \Big(\pi - \theta_{\boldsymbol{w}, \boldsymbol{w}^*} \Big) \boldsymbol{w}^* \right] \end{aligned} \tag{2}$$

   and assume it vanishes.

   Denote $\theta \triangleq \theta_{\boldsymbol{w}, \boldsymbol{w}^*}$. If $\theta = 0$ then let $\boldsymbol{w} = \alpha \boldsymbol{w}^*$ for some $\alpha > 0$. It follows that

$$k + \frac{k^2 - k}{\pi} - \frac{k^2 - k}{\pi} \frac{1}{\alpha} - \frac{k}{\alpha} = 0$$

   or equivalently $\alpha = 1$, and thus $\boldsymbol{w} = \boldsymbol{w}^*$.

   If $\theta = \pi$ then $\|\boldsymbol{w}\| = \frac{k^2 - k}{k^2 + (\pi - 1)k} \|\boldsymbol{w}^*\|$ and thus $\boldsymbol{w} = -(\frac{k^2 - k}{k^2 + (\pi - 1)k}) \boldsymbol{w}^*$. By setting $\theta = \pi$ in the loss function, one can see that $\boldsymbol{w} = -(\frac{k^2 - k}{k^2 + (\pi - 1)k}) \boldsymbol{w}^*$ is a one-dimensional local minimum, whereas by fixing $\|\boldsymbol{w}\|$ and decreasing θ, the loss function decreases. It follows that $\boldsymbol{w} = -(\frac{k^2 - k}{k^2 + (\pi - 1)k}) \boldsymbol{w}^*$ is a saddle point. If $\theta \neq 0, \pi$ then $\boldsymbol{w}$ and $\boldsymbol{w}^*$ are linearly independent and thus $\frac{k}{\pi} \Big(\pi - \theta \Big) = 0$ which is a contradiction.

   It remains to show that $\boldsymbol{u} = -\gamma(k) \boldsymbol{w}^*$ where $\gamma(k) = \frac{k^2 - k}{k^2 + (\pi - 1)k}$ is a degenerate saddle point. We will show that the Hessian at $\mathbf{u}$ denoted by $\nabla^2 \ell(\boldsymbol{u})$, has only nonnegative eigenvalues and at least one zero eigenvalue. Let $\tilde{\ell}(\boldsymbol{w}) \triangleq \ell(\boldsymbol{w}, R\boldsymbol{w}^*)$, where the second entry denotes the ground truth weight vector and R is a rotation matrix. Denote by $f_{\mathbf{d}_1, \mathbf{d}_2}$ the second directional derivative of a function f in directions $\mathbf{d}_1$ and $\mathbf{d}_2$. Similarly to the proof of Lemma 3.2, since ℓ depends only on $\|\boldsymbol{w}\|$, $\|\boldsymbol{w}^*\|$ and $\theta_{\boldsymbol{w}, \boldsymbol{w}^*}$, we notice that

$$\ell_{\mathbf{d}_1, \mathbf{d}_2}(\boldsymbol{w}) = \tilde{\ell}_{R\mathbf{d}_1, R\mathbf{d}_2}(R\boldsymbol{w})$$

or equivalently

$$\mathbf{d}_1^T \nabla^2 \ell(\boldsymbol{w})\mathbf{d}_2 = (R\mathbf{d}_1)^T \nabla^2 \tilde{\ell}(R\boldsymbol{w})R\mathbf{d}_2 = \mathbf{d}_1^T R^T \nabla^2 \tilde{\ell}(R\boldsymbol{w})R\mathbf{d}_2$$

for any $\boldsymbol{w}$ and directions $\mathbf{d}_1$ and $\mathbf{d}_2$. It follows that

$$\nabla^2 \ell(\boldsymbol{w}) = R^T \nabla^2 \tilde{\ell}(R\boldsymbol{w})R$$

for all $\boldsymbol{w}$. Since R is an orthogonal matrix, we have that $\nabla^2 \ell(\boldsymbol{w})$ and $\nabla^2 \tilde{\ell}(R\boldsymbol{w})$ are similar matrices and thus have the same eigenvalues. Therefore, we can w.l.o.g. rotate $\boldsymbol{w}^*$ such that it will be on the w_1 axis.

By symmetry we have

$$\frac{\partial \ell}{\partial w_1 \partial w_i}(\boldsymbol{u}) = \frac{\partial \ell}{\partial w_1 \partial w_j}(\boldsymbol{u}),\ \frac{\partial \ell}{\partial w_i \partial w_1}(\boldsymbol{u}) = \frac{\partial \ell}{\partial w_j \partial w_1}(\boldsymbol{u})$$

and

$$\frac{\partial \ell}{\partial w_i^2}(\boldsymbol{u}) = \frac{\partial \ell}{\partial w_j^2}(\boldsymbol{u}),\ \frac{\partial \ell}{\partial w_i \partial w_j}(\boldsymbol{u}) = \frac{\partial \ell}{\partial w_s \partial w_t}(\boldsymbol{u})$$

for $i \neq j, s \neq t$ such that $i, j, s, t \neq 1$. It follows that we only need to consider second partial derivatives with respect to 3 axes w_1,w_2 and w_3. Denote $\boldsymbol{u}_\epsilon = (-\gamma(k), \epsilon, 0, ..., 0)$ and $\boldsymbol{w}^* = (1, 0, ..., 0)$ and $\beta(k) = \frac{k^2-k}{\pi}$ and note that $\gamma(k) = \frac{\beta(k)}{\beta(k)+k}$. Then by equation Eq. 2 we have

$$\begin{aligned}\frac{\partial \ell}{\partial w_2^2}(\boldsymbol{u}) &= \lim_{\epsilon \to 0} \frac{\nabla \ell(\boldsymbol{u}_\epsilon)_x - \nabla \ell(\boldsymbol{u})_x}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\frac{1}{k^2}\Big[\big(k+\beta(k)\big)\epsilon - \frac{k}{\pi}\left\|\boldsymbol{w}^*\right\| \frac{\epsilon}{\|\boldsymbol{u}_\epsilon\|} \sin\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*} - \beta(k)\left\|\boldsymbol{w}^*\right\| \frac{\epsilon}{\|\boldsymbol{u}_\epsilon\|}\Big]}{\epsilon} \\ &= \frac{1}{k^2}\big(k + \beta(k) - \frac{\beta(k)}{\gamma(k)}\big) = 0\end{aligned} \tag{3}$$

Furthermore,

$$\begin{aligned}\frac{\partial \ell}{\partial w_1 \partial w_2}(\boldsymbol{u}) &= \lim_{\epsilon \to 0} \frac{\nabla \ell(\boldsymbol{u}_\epsilon)_y - \nabla \ell(\boldsymbol{u})_y}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\frac{1}{k^2}\Big[-\big(k+\beta(k)\big)\gamma(k) + \frac{k}{\pi}\left\|\boldsymbol{w}^*\right\| \frac{\gamma(k)}{\|\boldsymbol{u}_\epsilon\|} \sin\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*} + \beta(k)\left\|\boldsymbol{w}^*\right\| \frac{\gamma(k)}{\|\boldsymbol{u}_\epsilon\|} - \frac{k}{\pi}(\pi - \theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*})\Big]}{\epsilon}\end{aligned} \tag{4}$$

where $\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*} = \arccos(\frac{-\gamma(k)}{\sqrt{\epsilon^2+\gamma^2(k)}})$.

By L'Hopital's rule we have

$$\begin{aligned}\frac{\partial \ell}{\partial w_1 \partial w_2}(\boldsymbol{u}) &= \lim_{\epsilon \to 0} -\frac{\gamma(k)\epsilon \sin\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*}}{\pi k \|\boldsymbol{u}_\epsilon\|^3} + \frac{\gamma(k)\cos\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*}\frac{\partial \theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*}}{\partial w_2}}{\pi k \left\|\boldsymbol{u}_\epsilon\right\|} - \frac{\beta(k)\gamma(k)\epsilon}{\left\|\boldsymbol{u}_\epsilon\right\|^3} + \frac{\frac{\partial \theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*}}{\partial w_2}}{\pi k} \\ &= \frac{1}{\pi k}\lim_{\epsilon \to 0} \frac{\partial \theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*}}{\partial w_2}\Big(\frac{\gamma(k)\cos\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*}}{\|\boldsymbol{u}_\epsilon\|} + 1\Big)\end{aligned} \tag{5}$$

Since

$$\frac{\partial \theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*}}{\partial w_2}(\boldsymbol{u}_\epsilon) = -\frac{1}{\frac{|\epsilon|}{\sqrt{\epsilon^2+\gamma^2(k)}}}\frac{\epsilon\gamma(k)}{(\epsilon^2+\gamma^2(k))^{\frac{3}{2}}} = -\frac{\epsilon\gamma(k)}{(\epsilon^2+\gamma^2(k))|\epsilon|}$$

it follows that

$$\begin{aligned} \left|\frac{\partial \ell}{\partial w_1 \partial w_2}(\boldsymbol{u})\right| &= \frac{1}{\pi k}\lim_{\epsilon\to 0}\left|\frac{\partial \theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*}}{\partial w_2}\right|\left|\frac{\gamma(k)\cos\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*}}{\|\boldsymbol{u}_\epsilon\|}+1\right| \\ &\leq \frac{1}{\gamma(k)\pi k}\lim_{\epsilon\to 0}\left|\frac{\gamma(k)\cos\theta_{\boldsymbol{u}_\epsilon,\boldsymbol{w}^*}}{\|\boldsymbol{u}_\epsilon\|}+1\right| = 0 \end{aligned} \tag{6}$$

and thus $\frac{\partial \ell}{\partial w_1 \partial w_2}(\boldsymbol{u}) = 0$.

Taking derivatives of the gradient with respect to w_1 is easier because the expressions in Eq. 2 that depend on $\theta_{\boldsymbol{w},\boldsymbol{w}^*}$ and $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$ are constant. Therefore,

$$\frac{\partial \ell}{\partial w_1^2}(\boldsymbol{u}) = \frac{k+\beta(k)}{k^2}$$

and

$$\frac{\partial \ell}{\partial w_2 \partial w_1}(\boldsymbol{u}) = 0$$

Finally let $\tilde{\boldsymbol{u}}_\epsilon = (0, -\gamma(k), \epsilon, 0, ..., 0)$ then it is easy to see that

$$\frac{\partial \ell}{\partial w_2 \partial w_3}(\boldsymbol{u}) = \lim_{\epsilon\to 0}\frac{\nabla \ell(\tilde{\boldsymbol{u}}_\epsilon)_{w_2} - \nabla \ell(\boldsymbol{u})_{w_2}}{\epsilon} = 0$$

.

Therefore, overall we see that $\nabla^2 \ell(\boldsymbol{u})$ is a diagonal matrix with zeros and $\frac{k+\beta(k)}{k^2} > 0$ on the diagonal, which proves our claim.

C.2 Proof of Theorem 5.2

For the following lemmas let $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \lambda \nabla \ell(\boldsymbol{w}_t)$, θ_t be the angle between $\boldsymbol{w}_t$ and $\boldsymbol{w}^*$ $(t \geq 0)$ and define $\tilde{\lambda} = \alpha(k)\lambda$ where $\alpha(k) = \frac{1}{k} + \frac{k^2-k}{\pi k^2}$. Note that $\alpha(k) \leq 1$ for all $k \geq 1$ The following lemma shows that for $\lambda < 1$, the angle between $\boldsymbol{w}_t$ and $\boldsymbol{w}^*$ decreases in each iteration.

Lemma C.1. *If $0 < \theta_t < \pi$ and $\lambda < 1$ then $\theta_{t+1} < \theta_t$.*

Proof. This follows from the fact that adding

$$-\frac{\lambda}{k^2}\left(k + \frac{k^2-k}{\pi} - \frac{k\,\|\boldsymbol{w}^*\|}{\pi\,\|\boldsymbol{w}_t\|}\sin\theta_t - \frac{k^2-k}{\pi}\frac{\|\boldsymbol{w}^*\|}{\|\boldsymbol{w}_t\|}\right)\boldsymbol{w}_t$$

to $\boldsymbol{w}_t$ does not change θ_t for $\lambda < 1$, since $\frac{k+\frac{k^2-k}{\pi}}{k^2} \leq 1$ for $k \geq 1$. In addition, adding $\frac{\lambda}{\pi k}\Big(\pi - \theta\Big)\boldsymbol{w}^*$ decreases θ_t. □

We will need the following two lemmas to establish a lower bound on $\|\boldsymbol{w}_t\|$.

Lemma C.2. *If $\frac{\pi}{2} < \theta_t < \pi$ then $\|\boldsymbol{w}_{t+1}\| \geq \frac{\sin\theta_t}{\sin\theta_{t+1}}\min\{\|\boldsymbol{w}_t\|, \frac{\|\boldsymbol{w}^*\|\sin\theta_t}{\alpha(k)\pi}\}$.*

Proof. Let

$$\boldsymbol{u}_t = \boldsymbol{w}_t - \frac{\lambda}{k^2}\left(k + \frac{k^2-k}{\pi} - \frac{k\left\|\boldsymbol{w}^*\right\|}{\pi\left\|\boldsymbol{w}_t\right\|}\sin\theta_t - \frac{k^2-k}{\pi}\frac{\left\|\boldsymbol{w}^*\right\|}{\left\|\boldsymbol{w}_t\right\|}\right)\boldsymbol{w}_t$$

Notice that if $\left\|\boldsymbol{w}_t\right\| \leq \frac{\left\|\boldsymbol{w}^*\right\|\sin\theta_t}{\alpha(k)\pi}$ then

$$\left\|\boldsymbol{u}_t\right\| = (1-\tilde{\lambda})\left\|\boldsymbol{w}_t\right\| + \frac{\lambda\left\|\boldsymbol{w}^*\right\|}{\pi k}\sin\theta_t + \frac{\lambda(k^2-k)\left\|\boldsymbol{w}^*\right\|}{\pi k^2} \geq$$

$$(1-\tilde{\lambda})\left\|\boldsymbol{w}_t\right\| + \frac{\lambda k\left\|\boldsymbol{w}^*\right\|\sin\theta_t}{\pi k^2} + \frac{\lambda(k^2-k)\left\|\boldsymbol{w}^*\right\|\sin\theta_t}{\pi k^2} =$$

$$(1-\tilde{\lambda})\left\|\boldsymbol{w}_t\right\| + \frac{\tilde{\lambda}\left\|\boldsymbol{w}^*\right\|\sin\theta_t}{\alpha(k)\pi} \geq \left\|\boldsymbol{w}_t\right\|$$

Similarly, if $\left\|\boldsymbol{w}_t\right\| \geq \frac{\left\|\boldsymbol{w}^*\right\|\sin\theta_t}{\alpha(k)\pi}$ then $\left\|\boldsymbol{u}_t\right\| \geq \frac{\left\|\boldsymbol{w}^*\right\|\sin\theta_t}{\alpha(k)\pi}$. Furthermore, by a simple geometric observation we see that $\left\|\boldsymbol{w}_{t+1}\right\|\cos(\theta_{t+1}-\frac{\pi}{2}) = \left\|\boldsymbol{u}_t\right\|\cos(\theta_t-\frac{\pi}{2})$ if $\theta_{t+1} > \frac{\pi}{2}$ and $\left\|\boldsymbol{w}_{t+1}\right\|\cos(\frac{\pi}{2}-\theta_{t+1}) = \left\|\boldsymbol{u}_t\right\|\cos(\theta_t-\frac{\pi}{2})$ if $\theta_{t+1} \leq \frac{\pi}{2}$. This is equivalent to $\left\|\boldsymbol{w}_{t+1}\right\| = \frac{\sin\theta_t}{\sin\theta_{t+1}}\left\|\boldsymbol{u}_t\right\|$. It follows that $\left\|\boldsymbol{w}_{t+1}\right\| \geq \frac{\sin\theta_t}{\sin\theta_{t+1}}\min\{\left\|\boldsymbol{w}_t\right\|, \frac{\left\|\boldsymbol{w}^*\right\|\sin\theta_t}{\alpha(k)\pi}\}$ as desired. □

Lemma C.3. *If* $0 < \theta_t \leq \frac{\pi}{2}$ *and* $0 < \lambda < \frac{1}{2}$ *then* $\left\|\boldsymbol{w}_{t+1}\right\| \geq \min\{\left\|\boldsymbol{w}_t\right\|, \frac{\left\|\boldsymbol{w}^*\right\|}{8}\}$

Proof. First assume that $k \geq 2$. Let $\boldsymbol{u}_t$ be as in Lemma C.2, then

$$\left\|\boldsymbol{u}_t\right\| \geq (1-\tilde{\lambda})\left\|\boldsymbol{w}_t\right\| + \frac{\tilde{\lambda}(k^2-k)\left\|\boldsymbol{w}^*\right\|}{\alpha(k)\pi k^2}$$

It follows that if $\left\|\boldsymbol{w}_t\right\| \geq \frac{(k^2-k)\left\|\boldsymbol{w}^*\right\|}{\alpha(k)\pi k^2} \geq \frac{\left\|\boldsymbol{w}^*\right\|}{2\pi}$ then $\left\|\boldsymbol{u}_t\right\| \geq \frac{\left\|\boldsymbol{w}^*\right\|}{2\pi}$. Otherwise if $\left\|\boldsymbol{w}_t\right\| \leq \frac{(k^2-k)\left\|\boldsymbol{w}^*\right\|}{\alpha(k)\pi k^2}$ then $\left\|\boldsymbol{u}_t\right\| \geq \left\|\boldsymbol{w}_t\right\|$. Since $\boldsymbol{w}_{t+1} = \boldsymbol{u}_t + \frac{\lambda}{\pi k}\Big(\pi - \theta\Big)\boldsymbol{w}^*$ and $0 < \theta_t \leq \frac{\pi}{2}$ we have $\left\|\boldsymbol{w}_{t+1}\right\| \geq \left\|\boldsymbol{u}_t\right\| \geq \min\{\frac{\left\|\boldsymbol{w}^*\right\|}{2\pi}, \left\|\boldsymbol{w}_t\right\|\}$.

Now let $k = 1$. Note that in this case $\tilde{\lambda} = \lambda$. First assume that $\theta_t < \frac{\pi}{3}$. If $\left\|\boldsymbol{w}_t\right\| \geq \frac{\left\|\boldsymbol{w}^*\right\|}{4}$ then, using the same notation as in Lemma C.2, $\left\|\boldsymbol{u}_t\right\| \geq (1-\lambda)\left\|\boldsymbol{w}_t\right\| + \frac{\lambda\left\|\boldsymbol{w}^*\right\|\sin\theta_t}{\pi} \geq \frac{\left\|\boldsymbol{w}_t\right\|}{2} \geq \frac{\left\|\boldsymbol{w}^*\right\|}{8}$. Since $\boldsymbol{w}_{t+1} = \boldsymbol{u}_t + \frac{\lambda}{\pi}\Big(\pi - \theta_t\Big)\boldsymbol{w}^*$ and $0 < \theta_t \leq \frac{\pi}{2}$ we have $\left\|\boldsymbol{w}_{t+1}\right\| \geq \left\|\boldsymbol{u}_t\right\| \geq \frac{\left\|\boldsymbol{w}^*\right\|}{8}$. If $\left\|\boldsymbol{w}_t\right\| < \frac{\left\|\boldsymbol{w}^*\right\|}{4}$ then by the facts $0 < \theta_t \leq \frac{\pi}{2}$ and $\cos\theta_t > \frac{1}{2}$ we get

$$\left\|\boldsymbol{w}_{t+1}\right\|^2 = \left\|\boldsymbol{u}_t\right\|^2 + 2\left\|\boldsymbol{u}_t\right\|\left\|\frac{\lambda}{\pi}\Big(\pi-\theta_t\Big)\boldsymbol{w}^*\right\|\cos\theta_t + \left\|\frac{\lambda}{\pi}\Big(\pi-\theta_t\Big)\boldsymbol{w}^*\right\|^2 \geq$$

$$(1-\lambda)^2\left\|\boldsymbol{w}_t\right\|^2 + \frac{(1-\lambda)\lambda}{2}\left\|\boldsymbol{w}_t\right\|\left\|\boldsymbol{w}^*\right\| + \frac{\lambda^2}{4}\left\|\boldsymbol{w}^*\right\|^2 \geq$$

$$(1-\lambda)^2\left\|\boldsymbol{w}_t\right\|^2 + 2(1-\lambda)\lambda\left\|\boldsymbol{w}_t\right\|^2 + 4\lambda^2\left\|\boldsymbol{w}_t\right\|^2 =$$

$$(1+3\lambda^2)\left\|\boldsymbol{w}_t\right\|^2 \geq \left\|\boldsymbol{w}_t\right\|^2$$

Finally, assume $\theta_t \geq \frac{\pi}{3}$. As in the proof of Lemma C.2, if $\left\|\boldsymbol{w}_t\right\| \geq \frac{\left\|\boldsymbol{w}^*\right\|\sin\theta_t}{\pi} \geq \frac{\sqrt{3}}{2}\frac{\left\|\boldsymbol{w}^*\right\|}{\pi}$ then $\left\|\boldsymbol{w}_{t+1}\right\| \geq \left\|\boldsymbol{u}_t\right\| \geq \frac{\sqrt{3}}{2}\frac{\left\|\boldsymbol{w}^*\right\|}{\pi}$. Otherwise, if $\left\|\boldsymbol{w}_t\right\| < \frac{\left\|\boldsymbol{w}^*\right\|\sin\theta_t}{\pi}$ then $\left\|\boldsymbol{w}_{t+1}\right\| \geq \left\|\boldsymbol{u}_t\right\| \geq \left\|\boldsymbol{w}_t\right\|$. This concludes our proof. □

We can now show that in each iteration $\left\|\boldsymbol{w}_t\right\|$ is bounded away from 0 by a constant.

Proposition C.4. *Assume GD is initialized at $\boldsymbol{w}_0$ such that $\theta_0 \neq \pi$ and runs for T iterations with learning rate $0 < \lambda < \frac{1}{2}$. Then for all $0 \leq t \leq T$,*

$$\|\boldsymbol{w}_t\| \geq \min\{\|\boldsymbol{w}_0\| \sin\theta_0, \frac{\|\boldsymbol{w}^*\| \sin^2\theta_0}{\alpha(k)\pi}, \frac{\|\boldsymbol{w}^*\|}{8}\}$$

Proof. Let $\theta_0 > \theta_1 > ... > \theta_T$ (by Lemma C.1). Let i be the last index such that $\theta_i > \frac{\pi}{2}$ (if such i does not exist let $i = -1$). Since $\sin\theta_j > \sin\theta_0$ for all $0 \leq j \leq i$, by applying Lemma C.2 at most $j+1$ times we have

$$\|\boldsymbol{w}_{j+1}\| \geq \min\{\|\boldsymbol{w}_0\| \sin\theta_0, \frac{\|\boldsymbol{w}^*\| \sin^2\theta_0}{\alpha(k)\pi}\}$$

for all $0 \leq j \leq i$.

Finally, by Lemma C.3 and the fact that $\theta_j \leq \frac{\pi}{2}$ for all $i < j \leq T$, we get

$$\|\boldsymbol{w}_j\| \geq \min\{\|\boldsymbol{w}_{i+1}\|, \frac{\|\boldsymbol{w}^*\|}{8}\}$$

for all $i+1 < j \leq T$, from which the claim follows. □

The following lemma shows that $\nabla\ell$ is Lipschitz continuous at points that are bounded away from 0.

Lemma C.5. *Assume $\|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\| \geq M$, $\boldsymbol{w}_1, \boldsymbol{w}_2$ and $\boldsymbol{w}^*$ are on the same two dimensional half-plane defined by $\boldsymbol{w}^*$, then*

$$\|\nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2)\| \leq L \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$$

for $L = 1 + \frac{3\|\boldsymbol{w}^*\|}{M}$.

Proof. Recall that by equality Eq. 1,

$$\frac{\partial g}{\partial \boldsymbol{w}}(\boldsymbol{w}, \boldsymbol{w}^*) = \frac{1}{2\pi} \|\boldsymbol{w}^*\| \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \sin\theta_{\boldsymbol{w},\boldsymbol{w}^*} + \frac{1}{2\pi}\Big(\pi - \theta_{\boldsymbol{w},\boldsymbol{w}^*}\Big)\boldsymbol{w}^*$$

Let θ_1 and θ_2 be the angles between $\boldsymbol{w}_1$,$\boldsymbol{w}^*$ and $\boldsymbol{w}_2$,$\boldsymbol{w}^*$, respectively. By the inequality $\frac{x_0 \sin x}{\sin x_0} \geq x$ for $0 \leq x \leq x_0 < \pi$ and since $\frac{|\theta_1 - \theta_2|}{2} \leq \frac{\pi}{2}$ we have

$$\frac{|\theta_1 - \theta_2|}{2} \leq \frac{\pi \sin\frac{|\theta_1 - \theta_2|}{2}}{2}$$

Furthermore $\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$ is minimized (for fixed angles θ_1 and θ_2) when $\|\boldsymbol{w}_1\| = \|\boldsymbol{w}_2\| = M$ and is equal to $2M \sin\frac{|\theta_1 - \theta_2|}{2}$. Thus, under our assumptions we have,

$$\frac{|\theta_1 - \theta_2|}{2} \leq \frac{\pi \sin\frac{|\theta_1 - \theta_2|}{2}}{2} \leq \frac{\pi \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|}{4M}$$

Thus we get

$$\left\|\frac{1}{2\pi}\Big(\pi - \theta_1\Big)\boldsymbol{w}^* - \frac{1}{2\pi}\Big(\pi - \theta_2\Big)\boldsymbol{w}^*\right\| \leq \frac{\|\boldsymbol{w}^*\|}{4M} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$$

For the first summand, we will first find the parameterization of a two dimensional vector of length $\sin\theta$ where θ is the angle between the vector and the positive x axis. Denote this vector by (a, b), then the following holds

$$a^2 + b^2 = \sin^2\theta$$

and

$$\frac{b}{a} = \tan\theta$$

The solution to these equations is $(a, b) = (\frac{\sin 2\theta}{2}, \sin^2\theta)$. Hence (here we use the fact that $\boldsymbol{w}_1$,$\boldsymbol{w}_2$ are on the same half-plane)

$$\left\| \frac{1}{2\pi} \|\boldsymbol{w}^*\| \frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|} \sin\theta_1 - \frac{1}{2\pi} \|\boldsymbol{w}^*\| \frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|} \sin\theta_2 \right\| =$$

$$\frac{1}{2\pi} \|\boldsymbol{w}^*\| \sqrt{\Big(\frac{\sin 2\theta_1}{2} - \frac{\sin 2\theta_2}{2}\Big)^2 + \Big(\sin^2\theta_1 - \sin^2\theta_2\Big)^2} \leq$$

$$\frac{1}{2\pi} \|\boldsymbol{w}^*\| \sqrt{(\theta_1 - \theta_2)^2 + 4(\theta_1 - \theta_2)^2} \leq$$

$$\frac{\sqrt{5}}{\pi} \|\boldsymbol{w}^*\| \frac{\pi \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|}{4M} = \frac{\sqrt{5}\,\|\boldsymbol{w}^*\|}{4M} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$$

where the first inequality follows from the fact that $|\sin x - \sin y| \leq |x - y|$ and the second inequality from previous results. In conclusion, we have

$$\left\| \frac{\partial g}{\partial \boldsymbol{w}}(\boldsymbol{w}_1, \boldsymbol{w}^*) - \frac{\partial g}{\partial \boldsymbol{w}}(\boldsymbol{w}_2, \boldsymbol{w}^*) \right\| \leq \frac{(\sqrt{5} + 1)\,\|\boldsymbol{w}^*\|}{4M} \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$$

Similarly, in order to show that the function $f(\boldsymbol{w}) = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$ is Lipschitz continuous, we parameterize the unit vector by $(\cos\theta, \sin\theta)$ where θ is the angle between the vector and the positive x axis. We now obtain

$$\left\| \frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|} - \frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|} \right\| = \sqrt{(\cos\theta_1 - \cos\theta_2)^2 + (\sin\theta_1 - \sin\theta_2)^2} \leq$$

$$\sqrt{2(\theta_1 - \theta_2)^2} \leq \frac{\pi \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|}{\sqrt{2}M}$$

Now we can conclude that

$$\|\nabla\ell(\boldsymbol{w}_1) - \nabla\ell(\boldsymbol{w}_2)\| \leq \big(\frac{1}{k} + \frac{k^2 - k}{\pi k^2}\big) \|\boldsymbol{w}_1 - \boldsymbol{w}_2\| + \frac{2}{k} \left\| \frac{\partial g}{\partial \boldsymbol{w}}(\boldsymbol{w}_1, \boldsymbol{w}^*) - \frac{\partial g}{\partial \boldsymbol{w}}(\boldsymbol{w}_2, \boldsymbol{w}^*) \right\| +$$

$$\Big(\frac{(k^2 - k)\,\|\boldsymbol{w}^*\|}{\pi k^2}\Big) \left\| \frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|} - \frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|} \right\| \leq \Big(\frac{1}{k} + \frac{k^2 - k}{\pi k^2} + \frac{(k^2 - k)\,\|\boldsymbol{w}^*\|}{\sqrt{2}Mk^2} + \frac{(\sqrt{5} + 1)\,\|\boldsymbol{w}^*\|}{2Mk}\Big) \|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \leq$$

$$1 + \frac{\|\boldsymbol{w}^*\|}{\sqrt{2}M} + \frac{(\sqrt{5} + 1)\,\|\boldsymbol{w}^*\|}{2M} \leq 1 + \frac{3\,\|\boldsymbol{w}^*\|}{M}$$

□

Given that ℓ is Lipschitz continuous we can now follow standard optimization analysis (Nesterov (2004)) to show that $\lim_{t\to\infty} \|\nabla\ell(\mathbf{w}_t)\| = 0$.

Proposition C.6. *Assume GD is initialized at* $\mathbf{w}_0$ *such that* $\theta_0 \neq \pi$ *and runs with a constant learning rate* $0 < \lambda < \min\{\frac{2}{L}, \frac{1}{2}\}$ *where* $L = \tilde{O}(1)$. *Then for all* T

$$\sum_{t=0}^{T} \|\nabla\ell(\mathbf{w}_t)\|^2 \leq \frac{1}{\lambda(1 - \frac{\lambda}{2}L)} \ell(\mathbf{w}_0)$$

Proof. We will need the following lemma

Lemma C.7. *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be a continuously differentiable function on a set* $D \subseteq \mathbb{R}^n$ *and* $x, y \in D$ *such that for all* $0 \leq \tau \leq 1$, $x + \tau(y - x) \in D$ *and* $\|\nabla f(x + \tau(y - x)) - \nabla f(x)\| \leq L\,\|x - y\|$. *Then we have*

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \frac{L}{2} \|x - y\|^2$$

Proof. The proof exactly follows the proof of Lemma 1.2.3 in Nesterov (2004) and note that the proof only requires Lipschitz continuity of the gradient on the set $S = \{x + \tau(y - x) \mid 0 \leq \tau \leq 1\}$ and that $S \subseteq D$. □

By Proposition C.4, for all t, $\|\boldsymbol{w}_t\| \geq M'$ where

$$M' = \min\{\|\boldsymbol{w}_0\| \sin\theta_0, \frac{\|\boldsymbol{w}^*\| \sin^2\theta_0}{\alpha(k)\pi}, \frac{\|\boldsymbol{w}^*\|}{8}\}$$

. Furthermore, by a simple geometric observation we have

$$\min_{0\leq\tau\leq1, \|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\| \geq M', \arccos\left(\frac{\boldsymbol{w}_1 \cdot \boldsymbol{w}_2}{\|\boldsymbol{w}_1\|\|\boldsymbol{w}_2\|}\right)=\theta} \|\tau\boldsymbol{w}_1 + (1-\tau)\boldsymbol{w}_2\| = M' \cos\frac{\theta}{2}$$

.

It follows by Lemma C.5 that for any t and $\boldsymbol{x}_1, \boldsymbol{x}_2 \in S_t \triangleq \{\boldsymbol{w}_t + \tau(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t) \mid 0 \leq \tau \leq 1\}$,

$$\|\nabla\ell(\boldsymbol{x}_1) - \nabla\ell(\boldsymbol{x}_2)\| \leq L \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$$

where $L = 1 + \frac{3\|\boldsymbol{w}^*\|}{M}$ and $M = M' \cos\frac{\theta_0}{2}$ (Note that $\cos\frac{\theta_t - \theta_{t+1}}{2} \geq \cos\frac{\theta_0}{2}$ for all t by Lemma C.1).

Hence by Lemma C.7, for any t we have

$$\ell(\boldsymbol{w}_{t+1}) \leq \ell(\boldsymbol{w}_t) + \langle\nabla\ell(\boldsymbol{w}_t), \boldsymbol{w}_{t+1} - \boldsymbol{w}_t\rangle + \frac{L}{2}\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|^2 =$$

$$\ell(\boldsymbol{w}_t) - \lambda(1 - \frac{\lambda}{2}L)\|\nabla\ell(\boldsymbol{w}_t)\|^2$$

which implies that

$$\sum_{t=0}^{T} \|\nabla\ell(\boldsymbol{w}_t)\|^2 \leq \frac{1}{\lambda(1 - \frac{\lambda}{2}L)}\Big(\ell(\boldsymbol{w}_0) - \ell(\boldsymbol{w}_T)\Big) \leq \frac{1}{\lambda(1 - \frac{\lambda}{2}L)}\ell(\boldsymbol{w}_0)$$

□

We are now ready to prove the theorem.

Proof of Theorem 5.2. First, we observe that for a randomly initialized point $\boldsymbol{w}_0$, $0 \leq \theta_0 \leq \pi(1-\delta)$ with probability $1-\delta$. Hence by Proposition C.6 we have for $L = 1 + \frac{3\|\boldsymbol{w}^*\|}{M}$ where $M = \min\{\sin(\pi(1-\delta)), \frac{\sin^2(\pi(1-\delta))}{\alpha(k)\pi}, \frac{1}{8}\}\cos(\frac{\pi(1-\delta)}{2})$ and $\alpha(k) = k + \frac{k^2-k}{\pi}$, and for $\lambda = \frac{1}{L}$ (we assume w.l.o.g. that $L > 2$),

$$\sum_{t=0}^{T} \|\nabla\ell(\boldsymbol{w}_t)\|^2 \leq \frac{1}{\lambda(1 - \frac{\lambda}{2}L)}\ell(\boldsymbol{w}_0) = 2L\ell(\boldsymbol{w}_0) \leq \frac{4L}{k^2}(\frac{k}{2} + \frac{k^2-k}{2\pi})$$

Therefore,

$$\min_{0\leq t\leq T}\{\|\nabla\ell(\boldsymbol{w}_t)\|^2\} \leq \frac{\frac{4L}{k^2}\left(\frac{k}{2} + \frac{k^2-k}{2\pi}\right)}{T}$$

It follows that gradient descent reaches a point $\boldsymbol{w}_t$ such that $\|\nabla\ell(\boldsymbol{w}_t)\| < \epsilon$ after T iterations where

$$T > \frac{\left(\frac{4L}{k^2}\left(\frac{k}{2} + \frac{k^2-k}{2\pi}\right)\right)^2}{\epsilon^2}$$

We will now show that if $\|\nabla\ell(\boldsymbol{w}_t)\| < \epsilon$ then $\boldsymbol{w}_t$ is $O(\sqrt{\epsilon})$-close to the global minimum $\boldsymbol{w}^*$. First note that if $\frac{\pi}{2} \leq \theta_t \leq \pi(1-\delta)$ then a vector of the form $\mathbf{v} = \alpha\mathbf{w}^* + \beta\mathbf{w}$ where $\alpha \geq 0$ is of minimal norm equal to $\alpha \sin(\pi - \theta_t)\left\|\mathbf{w}^*\right\|$ when it is perpendicular to $\mathbf{w}$. Since the gradient is a vector of this form, we have $\|\nabla \ell(\boldsymbol{w}_t)\| > \frac{\pi\delta\|\boldsymbol{w}^*\|\sin\pi\delta}{\pi k} \geq \frac{\delta \sin \pi\delta}{k} \geq \epsilon$. Hence, from now on we assume that $0 \leq \theta_t < \frac{\pi}{2}$.

Similarly to the previous argument, we have

$$\epsilon > \|\nabla \ell(\boldsymbol{w}_t)\| > \frac{\|\boldsymbol{w}^*\|\,(\pi - \frac{\pi}{2})\sin\theta_t}{\pi k} \geq \frac{\sin\theta_t}{2k}$$

Hence, $\theta_t < \arcsin(2k\epsilon) = O(\epsilon)$. It follows by the triangle inequality that

$$k^2\epsilon > k^2\left\|\nabla \ell(\boldsymbol{w}_t)\right\| = \left\|\left(k + \frac{k^2-k}{\pi} - \frac{k\left\|\boldsymbol{w}^*\right\|}{\pi\left\|\boldsymbol{w}_t\right\|}\sin\theta_t - \frac{k^2-k}{\pi}\frac{\left\|\boldsymbol{w}^*\right\|}{\left\|\boldsymbol{w}_t\right\|}\right)\boldsymbol{w}_t - \frac{k(\pi-\theta_t)}{\pi}\boldsymbol{w}^*\right\| \geq$$

$$\left\|\left(k + \frac{k^2-k}{\pi} - \frac{k^2-k}{\pi}\frac{\left\|\boldsymbol{w}^*\right\|}{\left\|\boldsymbol{w}_t\right\|}\right)\boldsymbol{w}_t - k\boldsymbol{w}^*\right\| - \frac{k\left\|\boldsymbol{w}^*\right\|}{\pi}\sin\theta_t - \frac{k\theta_t\left\|\boldsymbol{w}^*\right\|}{\pi} \geq$$

$$\left\|\left(k + \frac{k^2-k}{\pi} - \frac{k^2-k}{\pi}\frac{\left\|\boldsymbol{w}^*\right\|}{\left\|\boldsymbol{w}_t\right\|}\right)\boldsymbol{w}_t - \frac{k\left\|\boldsymbol{w}^*\right\|}{\left\|\boldsymbol{w}_t\right\|}\boldsymbol{w}_t\right\| -$$

$$\left\|k\boldsymbol{w}^* - \frac{k\left\|\boldsymbol{w}^*\right\|}{\left\|\boldsymbol{w}_t\right\|}\boldsymbol{w}_t\right\| - \frac{k\left\|\boldsymbol{w}^*\right\|}{\pi}\sin\theta_t - \frac{k\theta_t\left\|\boldsymbol{w}^*\right\|}{\pi} \geq$$

$$\big(k + \frac{k^2-k}{\pi}\big)\left|\,\left\|\boldsymbol{w}_t\right\| - \left\|\boldsymbol{w}^*\right\|\,\right| - k\left\|\boldsymbol{w}^*\right\|\theta_t - \frac{k\left\|\boldsymbol{w}^*\right\|}{\pi}\sin\theta_t - \frac{k\theta_t\left\|\boldsymbol{w}^*\right\|}{\pi}$$

where the last inequality follows since the arc of a circle is larger than its corresponding segment.

Therefore we get $\left|\,\left\|\boldsymbol{w}_t\right\| - \left\|\boldsymbol{w}^*\right\|\,\right| < O(\epsilon)$. By the bounds on θ_t and $\left|\,\left\|\boldsymbol{w}_t\right\| - \left\|\boldsymbol{w}^*\right\|\,\right|$ and the inequality $\cos x \geq 1 - x$ for $x \geq 0$, we can give an upper bound on $\|\mathbf{w}_t - \mathbf{w}^*\|$:

$$\left\|\boldsymbol{w}_t - \boldsymbol{w}^*\right\|^2 = \left\|\boldsymbol{w}_t\right\|^2 - 2\left\|\boldsymbol{w}_t\right\|\left\|\boldsymbol{w}^*\right\|\cos\theta_t + \left\|\boldsymbol{w}^*\right\|^2 =$$

$$\left\|\boldsymbol{w}_t\right\|(\left\|\boldsymbol{w}_t\right\| - \left\|\boldsymbol{w}^*\right\|\cos\theta_t) + \left\|\boldsymbol{w}^*\right\|(\left\|\boldsymbol{w}^*\right\| - \left\|\boldsymbol{w}_t\right\|\cos\theta_t) \leq$$

$$(\left\|\boldsymbol{w}^*\right\| + O(\epsilon))(O(\epsilon) + \theta_t\left\|\boldsymbol{w}^*\right\|) + \left\|\boldsymbol{w}^*\right\|(O(\epsilon^2) + \theta_t\left\|\boldsymbol{w}^*\right\|) = O(\epsilon)$$

Finally, to prove the claim it suffices to show that $\ell(\mathbf{w}) \leq d\|\mathbf{w} - \mathbf{w}^*\|^2$. Denote the input vector $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)$ where $\mathbf{x}_i \in \mathbb{R}^m$ for all $1 \leq i \leq k$. Then we get

$$\begin{aligned}
\ell(\mathbf{w}) &= \mathbb{E}_{\mathbf{x}}\Big[\frac{\sum_{i=1}^k \sigma(\mathbf{w}^T\mathbf{x}_i)}{k} - \frac{\sum_{i=1}^k \sigma(\mathbf{w}^{*T}\mathbf{x}_i)}{k}\Big]^2 \\
&\leq \mathbb{E}_{\mathbf{x}}\Big[\frac{\sum_{i=1}^k |\sigma(\mathbf{w}^T\mathbf{x}_i) - \sigma(\mathbf{w}^{*T}\mathbf{x}_i)|}{k}\Big]^2 \\
&\leq \mathbb{E}_{\mathbf{x}}\Big[\frac{\sum_{i=1}^k |\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^{*T}\mathbf{x}_i|}{k}\Big]^2 \\
&\leq \mathbb{E}_{\mathbf{x}}\Big[\frac{\sum_{i=1}^k \left\|\mathbf{w} - \mathbf{w}^*\right\|\left\|\mathbf{x}_i\right\|}{k}\Big]^2 \\
&\leq \|\mathbf{w} - \mathbf{w}^*\|^2\mathbb{E}_{\mathbf{x}}\|\mathbf{x}\|^2 \\
&= d\|\mathbf{w} - \mathbf{w}^*\|^2
\end{aligned} \tag{7}$$

where the second inequality follows from Lipschitz continuity of σ, the third inequality from the Cauchy-Schwarz inequality and the last equality since $\|\mathbf{x}\|^2$ follows a chi-squared distribution with d degrees of freedom.

□

D Missing Proofs for Section 7.1

D.1 Proof of Proposition 7.1

Define $\boldsymbol{w}_p = (w_2, w_1)$, $\boldsymbol{w}^*_{p_1} = (0, -w^*)$ and $\boldsymbol{w}^*_{p_2} = (w^*, 0)$. We first prove the following lemma.

Lemma D.1. *Let l be defined as in Eq. 16. Then*

$$\begin{aligned}\nabla l(\boldsymbol{w}) = \frac{1}{k^2}\Bigg[&\big(k + \frac{k^2-3k+2}{\pi}\big)\boldsymbol{w} + \frac{2(k-1)\sin\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l}}{\pi}\boldsymbol{w} \\ &+ \frac{(k-1)(\pi - \theta_{\boldsymbol{w}_r,\boldsymbol{w}_l})}{\pi}\boldsymbol{w}_p - \frac{(k^2-3k+2)\left\|\boldsymbol{w}^*\right\|}{\pi\left\|\boldsymbol{w}\right\|}\boldsymbol{w} \\ &- \frac{k\left\|\boldsymbol{w}^*\right\|\sin\theta_{\boldsymbol{w},\boldsymbol{w}^*}}{\pi\left\|\boldsymbol{w}\right\|}\boldsymbol{w} - \frac{k(\pi - \theta_{\boldsymbol{w},\boldsymbol{w}^*})}{\pi}\boldsymbol{w}^* \\ &- \frac{(k-1)\sin\theta_{\boldsymbol{w}_l,\boldsymbol{w}^*_r}\left\|\boldsymbol{w}^*\right\|}{\pi\left\|\boldsymbol{w}\right\|}\boldsymbol{w} - \frac{(k-1)(\pi - \theta_{\boldsymbol{w}_l,\boldsymbol{w}^*_r})}{\pi}\boldsymbol{w}^*_{p_2} \\ &- \frac{(k-1)\sin\theta_{\boldsymbol{w}_r,\boldsymbol{w}^*_l}\left\|\boldsymbol{w}^*\right\|}{\pi\left\|\boldsymbol{w}\right\|}\boldsymbol{w} - \frac{(k-1)(\pi - \theta_{\boldsymbol{w}_r,\boldsymbol{w}^*_l})}{\pi}\boldsymbol{w}^*_{p_1}\Bigg]\end{aligned}$$

Proof. The gradient does not follow immediately from Lemma 3.2 because the loss has expressions with of the function g but with different dependencies on the parameters in A. We will only calculate $\frac{\partial g(\boldsymbol{w}_r,\boldsymbol{w}_l)}{\partial \boldsymbol{w}}$, the other expressions are calculated in the same manner.

Recall that

$$g(\boldsymbol{w}_r,\boldsymbol{w}_l) = \frac{1}{2\pi}\left\|\boldsymbol{w}\right\|^2(\sin\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l} + (\pi - \theta_{\boldsymbol{w}_r,\boldsymbol{w}_l})\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l})$$

It follows that

$$\frac{\partial g(\boldsymbol{w}_r,\boldsymbol{w}_l)}{\partial \boldsymbol{w}} = \frac{1}{\pi}(\sin\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l} + (\pi - \theta_{\boldsymbol{w}_r,\boldsymbol{w}_l})\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l})\boldsymbol{w} + \frac{1}{2\pi}\left\|\boldsymbol{w}\right\|^2(\pi - \theta_{\boldsymbol{w}_r,\boldsymbol{w}_l})\frac{\partial\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l}}{\partial \boldsymbol{w}} \tag{8}$$

Let $\boldsymbol{w} = (w_1, w_2)$ then $\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l} = \frac{w_1 w_2}{w_1^2 + w_2^2}$. Then,

$$\frac{\partial\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l}}{\partial w_1} = \frac{w_2(w_1^2+w_2^2) - 2w_1^2 w_2}{(w_1^2+w_2^2)^2} = \frac{w_2}{\left\|\boldsymbol{w}\right\|^2} - \frac{2w_1\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l}}{\left\|\boldsymbol{w}\right\|^2}$$

and

$$\frac{\partial\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l}}{\partial w_2} = \frac{w_1(w_1^2+w_2^2) - 2w_2^2 w_1}{(w_1^2+w_2^2)^2} = \frac{w_1}{\left\|\boldsymbol{w}\right\|^2} - \frac{2w_2\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l}}{\left\|\boldsymbol{w}\right\|^2}$$

or equivalently $\frac{\partial\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l}}{\partial \boldsymbol{w}} = \frac{\boldsymbol{w}_p}{\left\|\boldsymbol{w}\right\|^2} - \frac{2\boldsymbol{w}\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l}}{\left\|\boldsymbol{w}\right\|^2}$. It follows that

$$\frac{\partial g(\boldsymbol{w}_r,\boldsymbol{w}_l)}{\partial \boldsymbol{w}} = \frac{\sin\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l}\boldsymbol{w}}{\pi} + \frac{(\pi - \theta_{\boldsymbol{w}_l,\boldsymbol{w}_r})}{2\pi}\boldsymbol{w}_p$$

□

We will prove that $\boldsymbol{w}_{t+1} \neq 0$ and that it is in the interior of the fourth quadrant. Denote $\boldsymbol{w} = \boldsymbol{w}_t$ and $\nabla l(\boldsymbol{w}) = \frac{1}{k^2}\big(B_1(\boldsymbol{w}) + B_2(\boldsymbol{w}) + B_3(\boldsymbol{w})\big)$ where

$$B_1(\boldsymbol{w}) = \big(k + \frac{k^2-3k+2}{\pi}\big)\boldsymbol{w} + \frac{2(k-1)\sin\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l}}{\pi}\boldsymbol{w} -$$

$$\frac{(k^2-3k+2)\left\|\boldsymbol{w}^*\right\|}{\pi\left\|\boldsymbol{w}\right\|}\boldsymbol{w} - \frac{k\left\|\boldsymbol{w}^*\right\|\sin\theta_{\boldsymbol{w},\boldsymbol{w}^*}}{\pi\left\|\boldsymbol{w}\right\|}\boldsymbol{w} - \frac{(k-1)\sin\theta_{\boldsymbol{w}_l,\boldsymbol{w}^*_r}\left\|\boldsymbol{w}^*\right\|}{\pi\left\|\boldsymbol{w}\right\|}\boldsymbol{w} - \frac{(k-1)\sin\theta_{\boldsymbol{w}_r,\boldsymbol{w}^*_l}\left\|\boldsymbol{w}^*\right\|}{\pi\left\|\boldsymbol{w}\right\|}\boldsymbol{w}$$

$$B_2(\boldsymbol{w}) = \frac{(k-1)(\pi - \theta_{\boldsymbol{w}_r,\boldsymbol{w}_l})}{\pi}\boldsymbol{w}_p$$

and

$$B_3(\boldsymbol{w}) = -\frac{k(\pi - \theta_{\boldsymbol{w},\boldsymbol{w}^*})}{\pi}\boldsymbol{w}^* - \frac{(k-1)(\pi - \theta_{\boldsymbol{w}_l,\boldsymbol{w}_r^*})}{\pi}\boldsymbol{w}_{p_2}^* - \frac{(k-1)(\pi - \theta_{\boldsymbol{w}_r,\boldsymbol{w}_l^*})}{\pi}\boldsymbol{w}_{p_1}^*$$

Let $\boldsymbol{w} = (w, -mw)$ for $w, m \geq 0$. Straightforward calculation shows that $\cos\theta_{\boldsymbol{w}_l,\boldsymbol{w}_r^*} = \frac{1}{\sqrt{2(1+m^2)}}$ and $\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l^*} = \frac{m}{\sqrt{2(m^2+1)}}$. Hence $\frac{\pi}{4} \leq \theta_{\boldsymbol{w}_l,\boldsymbol{w}_r^*}, \theta_{\boldsymbol{w}_r,\boldsymbol{w}_l^*} \leq \frac{\pi}{2}$. Since $\boldsymbol{w}$ is in the fourth quadrant we also have $\frac{3\pi}{4} \leq \theta_{\boldsymbol{w},\boldsymbol{w}^*} \leq \pi$. Therefore, adding $-\lambda B_3(\boldsymbol{w})$ can only increase $\|\boldsymbol{w}\|$. This follows since in the worst case (the least possible increase of $\|\boldsymbol{w}\|$)

$$-B_3(\boldsymbol{w}) = \frac{k}{4}\boldsymbol{w}^* + \frac{k-1}{2}\boldsymbol{w}_{p_2}^* + \frac{k-1}{2}\boldsymbol{w}_{p_1}^* = (\frac{k-2}{4}w^*, -\frac{k-2}{4}w^*)$$

which is in the fourth quadrant for $k \geq 2$. In addition, since $-\boldsymbol{w}_p$ is in the fourth quadrant then adding $-\lambda B_2(\boldsymbol{w})$ increases $\|\boldsymbol{w}\|$.

If $\|\boldsymbol{w}\| < \frac{\|\boldsymbol{w}^*\|}{16}$ then $-B_1(\boldsymbol{w})$ points in the direction of $\boldsymbol{w}$ since in this case $-B_1(\boldsymbol{w}) = \alpha\boldsymbol{w}$ where

$$\alpha \geq \Big(\frac{k^2-3k+2}{\pi} + \frac{(k-1)}{\pi} - \frac{k-1}{8\pi} - \frac{k^2-3k+2}{16\pi} - \frac{k}{16}\Big)\|\boldsymbol{w}^*\| > 0$$

for $k \geq 2$. If $-B_1(\boldsymbol{w})$ points in the direction of $-\boldsymbol{w}$ then by the assumption that $\lambda \in (0, \frac{1}{3})$ we have $\|\lambda B_1(\boldsymbol{w})\| < \|\boldsymbol{w}\|$. Thus we can conclude that $\boldsymbol{w}_{t+1} \neq 0$.

Now, let $\boldsymbol{w} = (w_1, w_2)$, θ_t be the angle between $\boldsymbol{w} = \boldsymbol{w}_t$ and the positive x axis and first assume that $w_1 > -w_2$. In this case $-B_3(\boldsymbol{w})$ least increases (or even most decreases) θ_t when

$$-B_3(\boldsymbol{w}) = \frac{k}{4}\boldsymbol{w}^* + \frac{3(k-1)}{4}\boldsymbol{w}_{p_2}^* + \frac{k-1}{2}\boldsymbol{w}_{p_1}^* = \Big(\frac{2k-3}{4}w^*, \frac{2-k}{4}w^*\Big)$$

which is a vector in the fourth quadrant for $k \geq 2$. Otherwise, $-B_3(\boldsymbol{w})$ is a vector in the fourth quadrant as well. Note that we used the facts $\frac{\pi}{4} \leq \theta_{\boldsymbol{w}_l,\boldsymbol{w}_r^*}, \theta_{\boldsymbol{w}_r,\boldsymbol{w}_l^*} \leq \frac{\pi}{2}$ and $\frac{3\pi}{4} \leq \theta_{\boldsymbol{w},\boldsymbol{w}^*} \leq \pi$. Since $-\lambda B_1(\boldsymbol{w})$ does not change θ_t and $-\lambda B_2(\boldsymbol{w})$ increases θ_t but never to an angle greater than or equal to $\frac{\pi}{2}$, it follows that $0 < \theta_{t+1} < \frac{\pi}{2}$.

If $w_1 \leq -w_2$ then by defining all angles with respect to the negative y axis, we get the same argument as before. This shows that $\boldsymbol{w}_{t+1}$ is in the interior of the fourth quadrant, which concludes our proof.

D.2 Proof of Proposition 7.2

We will need the following auxiliary lemmas.

Lemma D.2. *Let $\boldsymbol{w}$ be in the fourth quadrant, then $g(\boldsymbol{w}_l, \boldsymbol{w}_r) \geq \frac{1}{2\pi}\big(\frac{\sqrt{3}}{2} - \frac{\pi}{6}\big)\|\boldsymbol{w}\|^2$.*

Proof. First note that the function $s(\theta) = \sin\theta + (\pi - \theta)\cos\theta$ is decreasing as a function of $\theta \in [0, \pi]$. Let $\boldsymbol{w} = (w, -mw)$ for $w, m \geq 0$. Straightforward calculation shows that $\cos\theta_{\boldsymbol{w}_l,\boldsymbol{w}_r} = -\frac{m}{m^2+1}$. As a function of $m \in [0, \infty)$, $\cos\theta_{\boldsymbol{w}_l,\boldsymbol{w}_r}$ is minimized for $m = 1$ with value $-\frac{1}{2}$, i.e., when $\theta(\boldsymbol{w}_l, \boldsymbol{w}_r) = \frac{2\pi}{3}$ and this is the largest angle possible. Thus $g(\boldsymbol{w}_l, \boldsymbol{w}_r) \geq \frac{1}{2\pi}s(\frac{2\pi}{3}))\|\boldsymbol{w}\|^2 = \frac{1}{2\pi}\big(\frac{\sqrt{3}}{2} - \frac{\pi}{6}\big)\|\boldsymbol{w}\|^2$. □

Lemma D.3. *Let*

$$f(\theta) = 2k\big(\sin(\frac{3\pi}{4} + \theta) + (\frac{\pi}{4} - \theta)\cos(\frac{3\pi}{4} + \theta)\big) +$$

$$\big(2k-2\big)\Big(\sqrt{1 - \frac{\cos\theta^2}{2}} + (\pi - \arccos\frac{\cos\theta}{\sqrt{2}})\frac{\cos\theta}{\sqrt{2}}\Big) + \big(2k-2\big)\Big(\sqrt{1 - \frac{\sin\theta^2}{2}} + (\pi - \arccos\frac{\sin\theta}{\sqrt{2}})\frac{\sin\theta}{\sqrt{2}}\Big)$$

, then in the interval $\theta \in [0, \frac{\pi}{4}]$, $f(\theta)$ is maximized at $\theta = \frac{\pi}{4}$ for all $k \geq 2$.

Proof. We will maximize the function $\frac{f(\theta)}{2(k-1)} = \frac{k}{k-1} f_1(\theta) + f_2(\theta) + f_3(\theta)$ where $f_1(\theta), f_2(\theta), f_3(\theta)$ correspond to the three summands in the expression of $f(\theta)$.

Since for $h(x) = \sqrt{1-x^2} + (\pi - \arccos(x))x$ we have $h'(x) = \pi - \arccos(x)$, it follows that $f_2'(\theta) = -(\pi - \arccos\frac{\cos\theta}{\sqrt{2}})\frac{\sin\theta}{\sqrt{2}}$, $f_3'(\theta) = (\pi - \arccos\frac{\sin\theta}{\sqrt{2}})\frac{\cos\theta}{\sqrt{2}}$ and $f_1'(\theta) = -(\frac{\pi}{4} - \theta)\sin(\frac{3\pi}{4} + \theta)$. It therefore suffices to show that

$$d_1(\theta) := (\pi - \arccos\frac{\sin\theta}{\sqrt{2}})\frac{\cos\theta}{\sqrt{2}} - (\pi - \arccos\frac{\cos\theta}{\sqrt{2}})\frac{\sin\theta}{\sqrt{2}} - \frac{k}{k-1}(\frac{\pi}{4} - \theta)\sin(\frac{3\pi}{4} + \theta) \geq 0$$

for $\theta \in [0, \frac{\pi}{4}]$.

By applying the inequalities $\arccos(x) \leq \frac{\pi}{2} - x$ for $x \in [0,1]$ and $\arccos(x) \geq \frac{\pi}{2} - x - \frac{1}{10}$ for $x \in [\frac{1}{2}, \frac{1}{\sqrt{2}}]$ we get $d_1(\theta) \geq d_2(\theta)$ where

$$d_2(\theta) = (\frac{\pi}{2} + \frac{\sin\theta}{\sqrt{2}})\frac{\cos\theta}{\sqrt{2}} - (\frac{\pi}{2} + \frac{\cos\theta}{\sqrt{2}} + \frac{1}{10})\frac{\sin\theta}{\sqrt{2}} - \frac{k}{k-1}(\frac{\pi}{4} - \theta)\sin(\frac{3\pi}{4} + \theta) =$$

$$\frac{\pi}{2\sqrt{2}}\cos\theta - (\frac{\pi}{2\sqrt{2}} + \frac{1}{10\sqrt{2}})\sin\theta - \frac{k}{k-1}(\frac{\pi}{4} - \theta)\sin(\frac{3\pi}{4} + \theta)$$

We notice that $d_2(0) \geq 0$ and $d_2(\frac{3}{4}) \geq 0$ for all $k \geq 2$. In addition,

$$d_2'(\theta) = -\frac{\pi}{2\sqrt{2}}\sin\theta - (\frac{\pi}{2\sqrt{2}} + \frac{1}{10\sqrt{2}})\cos\theta + \frac{k}{k-1}\sin(\frac{3\pi}{4} + \theta) - \frac{k}{k-1}(\frac{\pi}{4} - \theta)\cos(\frac{3\pi}{4} + \theta)$$

and $d_2'(0) > 0$ for all $k \geq 2$. It follows that in order to show that $d_2(\theta) \geq 0$ for $\theta \in [0, \frac{3}{4}]$ and $k \geq 2$, it suffices to show that $d_2''(\theta) \leq 0$ for $\theta \in [0, \frac{3}{4}]$ and $k \geq 2$. Indeed,

$$d_2''(\theta) = -\frac{\pi}{2\sqrt{2}}\cos\theta + (\frac{\pi}{2\sqrt{2}} + \frac{1}{10\sqrt{2}})\sin\theta + \frac{2k}{k-1}\cos(\frac{3\pi}{4} + \theta) + \frac{k}{k-1}(\frac{\pi}{4} - \theta)\sin(\frac{3\pi}{4} + \theta) \leq$$

$$(\frac{1}{10\sqrt{2}} + \frac{k}{k-1}\frac{\pi}{4})\max\{\sin\theta, \sin(\frac{3\pi}{4} + \theta)\} + \frac{2k}{k-1}\cos(\frac{3\pi}{4} + \theta) \leq 0$$

for all $\theta \in [0, \frac{3}{4}]$ and $k \geq 2$. Note that the first inequality follows since $\cos\theta \geq \sin\theta$ and the second since $\cos(\frac{3\pi}{4} + \theta) \geq \max\{\sin\theta, \sin(\frac{3\pi}{4} + \theta)\}$, both for $\theta \in [0, \frac{3}{4}]$. This shows that $d_1(\theta) \geq 0$ for $\theta \in [0, \frac{3}{4}]$.

Now assume that $\theta \in [\frac{3}{4}, \frac{\pi}{4}]$. Since $d_1(\frac{3}{4}) \geq 0$ and $d_1(\frac{\pi}{4}) \geq 0$, it suffices to prove that $d_1'(\theta) \leq 0$ for $\theta \in [\frac{3}{4}, \frac{\pi}{4}]$. Indeed, for all $\theta \in [\frac{3}{4}, \frac{\pi}{4}]$

$$d_1'(\theta) = -(\pi - \arccos\frac{\cos\theta}{\sqrt{2}})\frac{\cos\theta}{\sqrt{2}} - (\pi - \arccos\frac{\sin\theta}{\sqrt{2}})\frac{\sin\theta}{\sqrt{2}} +$$

$$\frac{\cos^2\theta}{2\sqrt{1 - \frac{\sin^2\theta}{2}}} + \frac{\sin^2\theta}{2\sqrt{1 - \frac{\cos^2\theta}{2}}} + \frac{k}{k-1}\sin(\frac{3\pi}{4} + \theta) - \frac{k}{k-1}(\frac{\pi}{4} - \theta)\cos(\frac{3\pi}{4} + \theta) \leq$$

$$-(\pi - \arccos\frac{\cos(\frac{\pi}{4})}{\sqrt{2}})\frac{\cos(\frac{\pi}{4})}{\sqrt{2}} - (\pi - \arccos\frac{\sin(\frac{3}{4})}{\sqrt{2}})\frac{\sin(\frac{3}{4})}{\sqrt{2}} +$$

$$\frac{\cos^2(\frac{3}{4})}{2\sqrt{1 - \frac{\sin^2(\frac{\pi}{4})}{2}}} + \frac{\sin^2(\frac{\pi}{4})}{2\sqrt{1 - \frac{\cos^2(\frac{3}{4})}{2}}} + 2\sin(\frac{3\pi}{4} + \frac{3}{4}) - 2(\frac{\pi}{4} - \frac{3}{4})\cos(\frac{3\pi}{4} + \frac{3}{4}) < 0$$

We conclude that $d_1(\theta) \geq 0$ for all $\theta \in [0, \frac{\pi}{4}]$ as desired.

□

Proof of Proposition 7.2. First assume that $w_1 \geq -w_2$. Let θ be the angle between $\boldsymbol{w}$ and the positive x axis. Then $\cos\theta = \frac{w_1}{\|\boldsymbol{w}\|}$ and $\tan\theta = -\frac{w_2}{w_1}$. Therefore we get

$$\cos\theta_{\boldsymbol{w}_l,\boldsymbol{w}_r^*} = \frac{w_1}{\|\boldsymbol{w}\|\sqrt{2}} = \frac{\cos\theta}{\sqrt{2}}$$

and

$$\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l^*} = \frac{-w_2}{\|\boldsymbol{w}\|\sqrt{2}} = \frac{\cos\theta\tan\theta}{\sqrt{2}} = \frac{\sin\theta}{\sqrt{2}}$$

We can rewrite $\ell(\boldsymbol{w})$ as

$$\ell(\boldsymbol{w}) = \frac{1}{k^2}\Bigg[\frac{k^2-3k+2}{2\pi}(\|\boldsymbol{w}\| - \|\boldsymbol{w}^*\|)^2 + \frac{k}{2}\|\boldsymbol{w}\|^2 + 2(k-1)g(\boldsymbol{w}_r,\boldsymbol{w}_l) -$$

$$\frac{\|\boldsymbol{w}\|\,\|\boldsymbol{w}^*\|}{2\pi}\Big(2k\big(\sin(\frac{3\pi}{4}+\theta) + (\frac{\pi}{4}-\theta)\cos(\frac{3\pi}{4}+\theta)\big)\Big) +$$

$$\big(2k-2\big)\big(\sqrt{1-\frac{\cos\theta^2}{2}} + (\pi - \arccos\frac{\cos\theta}{\sqrt{2}})\frac{\cos\theta}{\sqrt{2}}\big) + \big(2k-2\big)\big(\sqrt{1-\frac{\sin\theta^2}{2}} + (\pi - \arccos\frac{\sin\theta}{\sqrt{2}})\frac{\sin\theta}{\sqrt{2}}\big)\Big) +$$

$$\frac{k}{2}\|\boldsymbol{w}^*\|^2 + 2(k-1)g(\boldsymbol{w}_r^*,\boldsymbol{w}_l^*)\Bigg]$$

Hence by Lemma D.2 and Lemma D.3 we can lower bound $\ell(\boldsymbol{w})$ as follows

$$\ell(\boldsymbol{w}) \geq \frac{1}{k^2}\Bigg[\frac{k^2-3k+2}{2\pi}(\|\boldsymbol{w}\| - \|\boldsymbol{w}^*\|)^2 + \frac{k}{2}\|\boldsymbol{w}\|^2 + \frac{k-1}{\pi}(\frac{\sqrt{3}}{2} - \frac{\pi}{6})\|\boldsymbol{w}\|^2 -$$

$$\frac{(k-1)\,\|\boldsymbol{w}\|\,\|\boldsymbol{w}^*\|}{\pi}(\sqrt{3} + \frac{2\pi}{3}) + \frac{k}{2}\|\boldsymbol{w}^*\|^2 + \frac{k-1}{\pi}(\frac{\sqrt{3}}{2} - \frac{\pi}{6})\|\boldsymbol{w}^*\|^2\Bigg]$$

By setting $\|\boldsymbol{w}\| = \alpha\,\|\boldsymbol{w}^*\|$ we get

$$\frac{\ell(\boldsymbol{w})}{\|\boldsymbol{w}^*\|^2} \geq \frac{1}{k^2}\Bigg[\frac{k^2-3k+2}{2\pi}(\alpha-1)^2 + \frac{k}{2}\alpha^2 + \frac{k-1}{\pi}(\frac{\sqrt{3}}{2} - \frac{\pi}{6})\alpha^2 -$$

$$\frac{(k-1)}{\pi}(\sqrt{3} + \frac{2\pi}{3})\alpha + \frac{k}{2} + \frac{k-1}{\pi}(\frac{\sqrt{3}}{2} - \frac{\pi}{6})\Bigg]$$

Solving for α that minimizes the latter expression we obtain

$$\alpha^* = \frac{\frac{k^2-3k+2}{\pi} + \frac{(k-1)}{\pi}\big(\sqrt{3} + \frac{2\pi}{3}\big)}{k + \frac{k^2-3k+2}{\pi} + \frac{2(k-1)}{\pi}\big(\frac{\sqrt{3}}{2} - \frac{\pi}{6}\big)} = \frac{h(k)}{h(k)+1}$$

Plugging α^* back to the inequality we get

$$\ell(\boldsymbol{w}) \geq \frac{1}{k^2}\Big(\frac{h(k)+1}{2}(\alpha^*)^2 - h(k)\alpha^* + \frac{h(k)+1}{2}\Big)\|\boldsymbol{w}^*\|^2 = \frac{2h(k)+1}{k^2(2h(k)+2)}\|\boldsymbol{w}^*\|^2$$

and for $\tilde{\boldsymbol{w}} = -\alpha^*\boldsymbol{w}^*$ it holds that $\ell(\tilde{\boldsymbol{w}}) = \frac{2h(k)+1}{k^2(2h(k)+2)}\|\boldsymbol{w}^*\|^2$.

Finally, assume $w_1 \leq -w_2$. In this case, let θ be the angle between $\boldsymbol{w}$ and the negative y axis. Then $\cos\theta = \frac{-w_2}{\|\boldsymbol{w}\|}$ and $\tan\theta = -\frac{w_1}{w_2}$. Therefore

$$\cos\theta_{\boldsymbol{w}_l,\boldsymbol{w}_r^*} = \frac{w_1}{\|\boldsymbol{w}\|\sqrt{2}} = \frac{\cos\theta\tan\theta}{\sqrt{2}} = \frac{\sin\theta}{\sqrt{2}}$$

and

$$\cos\theta_{\boldsymbol{w}_r,\boldsymbol{w}_l^*} = \frac{-w_2}{\|\boldsymbol{w}\|\sqrt{2}} = \frac{\cos\theta}{\sqrt{2}}$$

Notice that from now on we get the same analysis as in the case where $w_1 \geq -w_2$, where we switch between expressions with $\boldsymbol{w}_l, \boldsymbol{w}_r^*$ and expressions with $\boldsymbol{w}_r, \boldsymbol{w}_l^*$. This concludes our proof. □

E Uniqueness of Global Minimum in the Population Risk

Without loss of generality we assume that the filter is of size 2 and the stride is 1. The proof of the general case follows the same lines. Assume that $\ell(\boldsymbol{w}) = 0$ and denote $\boldsymbol{w} = (w_1, w_2)$, $\boldsymbol{w}^* = (w_1^*, w_2^*)$. Recall that $\ell(\boldsymbol{w}) = \mathbb{E}_{\mathcal{G}}\left[(f(\boldsymbol{x};W) - f(\boldsymbol{x};W^*))^2\right]$ where $f(\boldsymbol{x};W) = \frac{1}{k}\sum_i \sigma\left(\boldsymbol{w}_i \cdot \boldsymbol{x}\right)$ and for all $1 \leq i \leq k$ $\boldsymbol{w}_i = (\boldsymbol{0}_{i-1}, \boldsymbol{w}, \boldsymbol{0}_{d-i-1})$. By equating $\ell(\boldsymbol{w})$ to 0 we get that $(f(\boldsymbol{x};W) - f(\boldsymbol{x};W^*))^2 = 0$ almost surely. Since $(f(\boldsymbol{x};W) - f(\boldsymbol{x};W^*))^2$ is a continuous function it follows that $f(\boldsymbol{x};W) - f(\boldsymbol{x};W^*) = 0$ for all $\boldsymbol{x}$. In particular this is true for $\boldsymbol{x}_1 = (x, 0, 0, ..., 0)$, $x \in \mathbb{R}$. Thus $\sigma\left(xw_1\right) = \sigma\left(xw_1^*\right)$ for all $x \in \mathbb{R}$ which implies that $w_1 = w_1^*$. The equality holds also for $\boldsymbol{x}_2 = (0, x, 0, ..., 0)$, $x \in \mathbb{R}$ which implies that $\sigma\left(xw_2\right) + \sigma\left(xw_1\right) = \sigma\left(xw_2^*\right) + \sigma\left(xw_1^*\right)$ for all $x \in \mathbb{R}$. By the previous result, we get $\sigma\left(xw_2\right) = \sigma\left(xw_2^*\right)$ for all $x \in \mathbb{R}$ and thus $w_2 = w_2^*$. We proved that $\boldsymbol{w} = \boldsymbol{w}^*$ and therefore $\boldsymbol{w}^*$ is the unique global minimum.

F Experimental Setup for Section 7.2

In our experiments we estimated the probability of convergence to the global minimum of a randomly initialized gradient descent for many different ground truths $\boldsymbol{w}^*$ of a convolutional neural network with overlapping filters. For each value of number of hidden neurons, filter size, stride length and ground truth distribution we randomly selected 30 different ground truths $\boldsymbol{w}^*$ with respect to the given distribution. We tested with all combinations of values given in Table 1.

Furthermore, for each combination of values of number of hidden neurons, filter size and stride length we tested with deterministic ground truths: ground truth with all entries equal to 1, all entries equal to -1 and with entries that form an increasing sequence from -1 to 1, -2 to 0 and 0 to 2 or decreasing sequence from 1 to -1, 0 to -2 and 2 to 0.

For each ground truth, we ran gradient descent 20 times and for each run we recorded whether it reached a point very close to the unique global minimum or it repeatedly (5000 consecutive iterations) incurred very low gradient values and stayed away from the global minimum. We then calculated the empirical probability $\hat{p} = \frac{\#\text{times reached global minimum}}{20}$. To compute the one-sided confidence interval we used the Wilson method (Brown et al. (2001)) which gives a lower bound

$$\frac{\hat{p} + \frac{z_\alpha^2}{2n} + z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_\alpha^2}{4k^2}}}{1 + \frac{z_\alpha^2}{n}} \tag{9}$$

where z_α is the Z-score with $\alpha = 0.05$ and in our experiments $n = 20$. Note that we initialized gradient descent inside a large hypercube such that outside the hypercube the gradient does not vanish (this can be easily proved after writing out the gradient for each setting).

For all ground truths we got $\hat{p} \geq 0.15$, i.e., for each ground truth we reached the global minimum at least 3 times. Hence the confidence interval lower bound Eq. 9 is greater than $\frac{1}{17}$ in all settings.

Table 1: Parameters values for experiments in Section 7.2

Number of hidden neurons	50,100
Filter size	2,8,16
stride length	$1,\min\{\frac{f}{4},1\},\min\{\frac{f}{2},1\}$ where f is the filter size (For instance, for $f = 16$ we used strides 1,4,8 and for $f = 2$ we used stride 1)
Ground truth distribution	The entries of the ground truth are i.i.d. uniform random variables over the interval $[a, b]$ where $(a, b) \in \{(-1, 1), (-2, 0), (0, 2)\}$

This suggests that with a few dozen repeated runs of a randomly initialized gradient descent, with high probability it will converge to the global minimum.

References

Brown, Lawrence D, Cai, T Tony, and DasGupta, Anirban. Interval estimation for a binomial proportion. *Statistical science*, pp. 101–117, 2001.

Nesterov, Yurii. Introductory lectures on convex optimization. pp. 22–29, 2004.