Appendix

A LEAST SQUARES ANALYSIS

A.1 Restated result

Recall that, after normalizing according to (9) and using the closed form solution for the optimal scaling factor $g^* := -\left(\mathbf{u}^\top \mathbf{w}\right)/\|\mathbf{w}\|_{\mathbf{S}}$, optimizing the ordinary least squares objective can be written as the following minimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} \left(\rho(\mathbf{w}) := -\frac{\mathbf{w}^\top \mathbf{u}\mathbf{u}^\top \mathbf{w}}{\mathbf{w}^\top \mathbf{S}\mathbf{w}}\right). \tag{12 revisited}$$

We consider optimizing the above objective by GD which takes iterative steps of the form

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla\rho(\mathbf{w}_t), \tag{23}$$

where

$$\nabla\rho(\mathbf{w}_t) = -2\left(\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{S}\mathbf{w}_t\right)/\mathbf{w}_t^\top \mathbf{S}\mathbf{w}_t. \tag{24}$$

Furthermore, we choose

$$\eta_t = \frac{\mathbf{w}_t^\top \mathbf{S}\mathbf{w}_t}{2L|\rho(\mathbf{w}_t)|}. \tag{25}$$

Our analysis relies on the (weak) data distribution assumption stated in A1. It is noteworthy that $\rho(\mathbf{w}) \leq 0, \forall \mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ holds under this assumption. In the next theorem, we establish a convergence rate of GD in terms of function value as well as gradient norm.

Theorem 1. *[Convergence rate on least squares] Suppose that the (weak) Assumption 1 on the data distribution holds. Consider the* GD *iterates* $\{\mathbf{w}_t\}_{t\in\mathbb{N}^+}$ *given in Eq.* (13) *with the stepsize* $\eta_t = \mathbf{w}_t^\top \mathbf{S}\mathbf{w}_t/(2L|\rho(\mathbf{w}_t)|)$ *and starting from* $\rho(\mathbf{w}_0) \neq 0$. *Then,*

$$\Delta\rho_t \leq \left(1 - \frac{\mu}{L}\right)^{2t} \Delta\rho_0, \tag{14}$$

where $\Delta\rho_t := \rho(\mathbf{w}_t) - \rho(\mathbf{w}^*)$. *Furthermore, the* $\mathbf{S}^{-1}$*-norm of the gradient* $\nabla\rho(\mathbf{w}_t)$ *relates to the suboptimality as*

$$\|\mathbf{w}_t\|_{\mathbf{S}}^2 \|\nabla\rho(\mathbf{w}_t)\|_{\mathbf{S}^{-1}}^2 / |4\rho(\mathbf{w}_t)| = \Delta\rho_t. \tag{15}$$

The proof of this result crucially relies on the insight that the minimization problem given in (11) resembles the problem of maximizing the generalized Rayleigh quotient which is commonly encountered in generalized eigenproblems. We will thus first review this area, where convergence rates are usually provided in terms of the angle of the current iterate with the maximizer, which is the principal eigenvector. Interestingly, this angle can be related to both, the current function value as well as the the norm of the current gradient. We will make use of these connections to prove the above Theorem in Section A.5. Although not necessarily needed for convex function, we introduce the gradient norm relation as we will later go on to prove a similar result for possibly non-convex functions in the learning halfspace setting (Theorem 4).

A.2 Background on eigenvalue problems

A.2.1 Rayleigh quotient

Optimizing the Rayleigh quotient is a classical non-convex optimization problem that is often encountered in eigenvector problems. Let $\mathbf{B} \in \mathbb{R}^{d\times d}$ be a symmetric matrix, then $\mathbf{w}^1 \in \mathbb{R}^d$ is the principal eigenvector of $\mathbf{B}$ if it

maximizes

$$q(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{B}\mathbf{w}}{\mathbf{w}^\top \mathbf{w}} \tag{26}$$

and $q(\mathbf{w})$ is called the Rayleigh quotient. Notably, this quotient satisfies the so-called Rayleigh inequality

$$\lambda_{\min}(\mathbf{B}) \leq q(\mathbf{w}) \leq \lambda_1(\mathbf{B}), \quad \forall \mathbf{w} \in \mathbb{R} \setminus \{\mathbf{0}\},$$

where $\lambda_{\min}(\mathbf{B})$ and $\lambda_1(\mathbf{B})$ are the smallest and largest eigenvalue of $\mathbf{B}$ respectively.

Maximizing $q(\mathbf{w})$ is a non-convex (strict-saddle) optimization problem, where the i-th critical point $\mathbf{v}_i$ constitutes the i-th eigenvector with corresponding eigenvalue $\lambda_i = q(\mathbf{w}_i)$ (see (Absil et al., 2009), Section 4.6.2 for details). It is known that optimizing $q(\mathbf{w})$ with GD - using an iteration-dependent stepsize - converges linearly to the principal eigenvector $\mathbf{v}_1$. The convergence analysis is based on the "minidimensional" method and yields the following result

$$\frac{\lambda_1 - q(\mathbf{w}_t)}{\cos^2 \angle(\mathbf{w}_t, \mathbf{v}_1)} \leq \left(1 - \frac{\lambda_1 - \lambda_2}{\lambda_1 - \lambda_{\min}}\right)^{2t} \frac{\lambda_1 - q(\mathbf{w}_0)}{\cos^2 \angle(\mathbf{w}_0, \mathbf{v}_1)} \tag{27}$$

under weak assumptions on $\mathbf{w}_0$. Details as well as the proof of this result can be found in (Knyazev and Shorokhodov, 1991).

A.2.2 Generalized rayleigh quotient

The reparametrized least squares objective (12), however, is not exactly equivalent to (26) because of the covariance matrix that appears in the denominator. As a matter of fact, our objective is a special instance of the generalized Rayleigh quotient

$$\tilde{\rho}(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{B}\mathbf{w}}{\mathbf{w}^\top \mathbf{A}\mathbf{w}}, \tag{28}$$

where $\mathbf{B}$ is defined as above and $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a symmetric positive definite matrix.

Maximizing (28) is a generalized eigenproblem in the sense that it solves the task of finding eigenvalues λ of the matrix pencil $(\mathbf{B}, \mathbf{A})$ for which $\det(\mathbf{B} - \lambda \mathbf{A}) = 0$, i.e. finding a vector $\mathbf{v}$ that obeys $\mathbf{B}\mathbf{v} = \lambda \mathbf{A}\mathbf{v}$. Again we have

$$\lambda_{\min}(\mathbf{B}, \mathbf{A}) \leq \tilde{\rho}(\mathbf{w}) \leq \lambda_1(\mathbf{B}, \mathbf{A}), \quad \forall \mathbf{w} \in \mathbb{R} \setminus \{0\}.$$

Among the rich literature on solving generalized symmetric eigenproblems, a GD convergence rate similar to (27) has been established in Theorem 6 of (Knyazev and Neymeyr, 2003), which yields

$$\frac{\lambda_1 - \rho(\mathbf{w}_{t+1})}{\rho(\mathbf{w}_{t+1}) - \lambda_2} \leq \left(1 - \frac{\lambda_1 - \lambda_2}{\lambda_1 - \lambda_{\min}}\right)^{2t} \frac{\lambda_1 - \rho(\mathbf{w}_t)}{\rho(\mathbf{w}_t) - \lambda_2},$$

again under weak assumptions on $\mathbf{w}_0$.

A.2.3 Our contribution

Since our setting in Eq. (12) is a minimization task, we note that for our specific choice of $\mathbf{A}$ and $\mathbf{B}$ we have $-\tilde{\rho}(\mathbf{w}) = \rho(\mathbf{w})$ and we recall the general result that then

$$\min \rho(\mathbf{w}) = -\max(\tilde{\rho}(\mathbf{w})).$$

More importantly, we here have a special case where the nominator of $\rho(\mathbf{w})$ has a particular low rank structure. In fact, $\mathbf{B} := \mathbf{u}\mathbf{u}^\top$ is a rank one matrix. Instead of directly invoking the convergence rate in (Knyazev and Neymeyr, 2003), this allows for a much simpler analysis of the convergence rate of GD on $\rho(\mathbf{w})$ since the rank one property yields a simpler representation of the relevant vectors. Furthermore, we establish a connection between suboptimality on function value and the $\mathbf{S}^{-1}$-norm of the gradient. As mentioned earlier, we need such a guarantee in our future analysis on learning halfspaces which is an instance of a (possibly) non-convex optimization problem.

A.3 Preliminaries

Notations Let $\mathbf{A}$ be a symmetric positive definite matrix. We introduce the following compact notations that will be used throughout the analysis.

- $\mathbf{A}$-inner product of $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$: $\langle \mathbf{w}, \mathbf{v} \rangle_{\mathbf{A}} = \mathbf{w}^\top \mathbf{A} \mathbf{v}$.
- $\mathbf{A}$-norm of $\mathbf{w} \in \mathbb{R}^d$: $\|\mathbf{w}\|_{\mathbf{A}} = \left(\mathbf{w}^\top \mathbf{A} \mathbf{w}\right)^{1/2}$.
- $\mathbf{A}$-angle between two vectors $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d n \{\mathbf{0}\}$:

$$\angle_{\mathbf{A}}(\mathbf{w}, \mathbf{v}) := \arccos\left(\frac{\langle \mathbf{w}, \mathbf{v} \rangle_{\mathbf{A}}}{\|\mathbf{w}\|_{\mathbf{A}} \|\mathbf{v}\|_{\mathbf{A}}}\right) \tag{29}$$

- $\mathbf{A}$-orthogonal projection $\hat{\mathbf{w}}$ of $\mathbf{w}$ to $\operatorname{span}\{\mathbf{v}\}$ for $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d n \{\mathbf{0}\}$:

$$\hat{\mathbf{w}} \in \operatorname{span}\{\mathbf{v}\} \quad \text{with} \quad \langle \mathbf{w} - \hat{\mathbf{w}}, \mathbf{v} \rangle_{\mathbf{A}} = 0 \tag{30}$$

- $\mathbf{A}$-spectral norm of a matrix $\mathbf{C} \in \mathbb{R}^{d \times d}$:

$$\|\mathbf{C}\|_{\mathbf{A}} := \max_{\mathbf{w} \in \mathbb{R}^d / \{\mathbf{0}\}} \frac{\|\mathbf{C}\mathbf{w}\|_{\mathbf{A}}}{\|\mathbf{w}\|_{\mathbf{A}}} \tag{31}$$

These notations allow us to make the analysis similar to the simple Rayleigh quotient case. For example, the denominator in (28) can now be written as $\|\mathbf{w}\|_{\mathbf{A}}^2$.

Properties We will use the following elementary properties of the induced terms defined above.

(P.1) $\sin^2 \angle_{\mathbf{A}}(\mathbf{w}, \mathbf{v}) = 1 - \cos^2 \angle_{\mathbf{A}}(\mathbf{w}, \mathbf{v})$

(P.2) If $\hat{\mathbf{w}}$ is the $\mathbf{A}$-orthogonal projection of $\mathbf{w}$ to $\operatorname{span}\{\mathbf{v}\}$, then it holds that

$$\cos^2 \angle_{\mathbf{A}}(\mathbf{w}, \mathbf{v}) = \frac{\|\hat{\mathbf{w}}\|_{\mathbf{A}}^2}{\|\mathbf{w}\|_{\mathbf{A}}^2}, \quad \sin \angle_{\mathbf{A}} \mathbf{w}\mathbf{v} = \frac{\|\mathbf{w} - \hat{\mathbf{w}}\|_{\mathbf{A}}}{\|\mathbf{w}\|_{\mathbf{A}}} \tag{32}$$

(P.3) The $\mathbf{A}$-spectral norm of a matrix can be written in the alternative form

$$\|\mathbf{C}\|_{\mathbf{A}} = \max_{\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{\|\mathbf{A}^{1/2}\mathbf{C}\mathbf{w}\|_2}{\|\mathbf{A}^{1/2}\mathbf{w}\|_2} \tag{33}$$

$$= \max_{\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{\| \left(\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{-1/2}\right) \mathbf{A}^{1/2}\mathbf{w}\|_2}{\|\mathbf{A}^{1/2}\mathbf{w}\|_2} \tag{34}$$

$$= \|\mathbf{A}^{1/2}\mathbf{C}\mathbf{A}^{-1/2}\|_2 \tag{35}$$

(P.4) Let $\mathbf{C} \in \mathbb{R}^{d \times d}$ be a square matrix and $\mathbf{w} \in \mathbb{R}^d$, then the following result holds due to the definition of $\mathbf{A}$-spectral norm

$$\|\mathbf{B}\mathbf{w}\|_{\mathbf{A}} \leq \|\mathbf{B}\|_{\mathbf{A}} \|\mathbf{w}\|_{\mathbf{A}} \tag{36}$$

A.4 Characterization of the LS minimizer

By setting the gradient of (10) to zero and recalling the convexity of f_{OLS} we immediately see that the minimizer of this objective is

$$\tilde{\mathbf{w}}^* := -\mathbf{S}^{-1}\mathbf{u}. \tag{37}$$

Indeed, one can easily verify that $\tilde{\mathbf{w}}^*$ is also an eigenvector of the matrix pair $(\mathbf{B}, \mathbf{S})$ since

$$\begin{aligned} \mathbf{B}\tilde{\mathbf{w}}^* &= \mathbf{u}\mathbf{u}^\top(-\mathbf{S}^{-1}\mathbf{u}) = -\|\mathbf{u}\|_{\mathbf{S}^{-1}}^2 \mathbf{u} \\ &= (\|\mathbf{u}\|_{\mathbf{S}^{-1}}^2)\mathbf{S}(-\mathbf{S}^{-1}\mathbf{u}) = \lambda_1 \mathbf{S}\tilde{\mathbf{w}}^* \end{aligned} \tag{38}$$

where $\lambda_1 := \|\mathbf{u}\|^2_{\mathbf{S}^{-1}}$ is the corresponding generalized eigenvalue. The associated eigenvector with λ_1 is

$$\mathbf{v}_1 := \tilde{\mathbf{w}}^*/\|\mathbf{u}\|_{\mathbf{S}^{-1}}. \tag{39}$$

Thereby we extend the normalized eigenvector to an $\mathbf{S}$-orthogonal basis $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d)$ of $\mathbb{R}^d$ such that

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle_{\mathbf{S}} = \begin{cases} 0, \ i \neq j \\ 1, \ i = j \end{cases} \tag{40}$$

holds for all i, j. Let $\mathbf{V}_2 := [\mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_d]$ be the matrix whose $(i-1)$-th column is $\mathbf{v}_i, i \in \{2, \ldots, d\}$. The matrix $\mathbf{B}$ is orthogonal to the matrix $\mathbf{V}_2$ since

$$\begin{aligned} \mathbf{B}\mathbf{V}_2 &= \mathbf{u}\mathbf{u}^\top \mathbf{V}_2 = \mathbf{u}\mathbf{u}^\top \mathbf{S}^{-1}\mathbf{S}\mathbf{V}_2 \\ &\overset{(39)}{=} \|\mathbf{S}^{-1}\mathbf{u}\|_{\mathbf{S}} \mathbf{u} \left(\underbrace{-\mathbf{v}_1^\top \mathbf{S}\mathbf{V}_2}_{\mathbf{0}} \right), \end{aligned} \tag{41}$$

which is a zero matrix due to the $\mathbf{A}$-orthogonality of the basis (see Eq. (40)). As a result, the columns of $\mathbf{V}_2$ are eigenvectors associated with a zero eigenvalue. Since $\mathbf{v}_1$ and $\mathbf{V}_2$ form an $\mathbf{S}$-orthonormal basis of $\mathbb{R}^d$ no further eigenvalues exist. We can conclude that any vector $\mathbf{w}^* \in \operatorname{span}\{\mathbf{v}_1\}$ is a minimizer of the reparametrized ordinary least squares problem as presented in (12) and the minimum value of ρ relates to the eigenvalue as

$$\lambda_1 = -\min_{\mathbf{w}} \rho(\mathbf{w}).$$

Spectral representation of suboptimality Our convergence analysis is based on the angle between the current iterate $\mathbf{w}_t$ and the leading eigenvector $\mathbf{v}_1$, for which we recall property (P.1). We can express $\mathbf{w} \in \mathbb{R}^d$ in the $\mathbf{S}$-orthogonal basis that we defined above:

$$\mathbf{w} = \alpha_1 \mathbf{v}_1 + \mathbf{V}_2 \boldsymbol{\alpha}_2, \quad \alpha_1 \in \mathbb{R}, \quad \boldsymbol{\alpha}_2 \in \mathbb{R}^{d-1} \tag{42}$$

and since $\alpha_1 \mathbf{v}_1$ is the $\mathbf{S}$-orthogonal projection of $\mathbf{w}$ to $\operatorname{span}\{\mathbf{v}_1\}$, the result of (P.2) implies

$$\cos^2 \angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1) = \frac{\|\alpha_1 \mathbf{v}_1\|^2_{\mathbf{S}}}{\|\mathbf{w}\|^2_{\mathbf{S}}} = \frac{\alpha_1^2}{\|\mathbf{w}\|^2_{\mathbf{S}}}. \tag{43}$$

Clearly this metric is zero for the optimal solution $\mathbf{v}_1$ and else bounded by one from above. To justify it is a proper choice, the next proposition proves that suboptimality on ρ, i.e. $\rho(\mathbf{w}) - \rho(\mathbf{v}_1)$, relates directly to this angle.

Proposition 1. *The suboptimality of* $\mathbf{w}$ *on* $\rho(\mathbf{w})$ *relates to* $\sin^2 \angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)$ *as*

$$\rho(\mathbf{w}) - \rho(\mathbf{v}_1) = \lambda_1 \sin^2 \angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1), \tag{44}$$

where $\rho(\mathbf{v}_1) = \lambda_1$. *This is equivalent to*

$$\rho(\mathbf{w}) = -\lambda_1 \cos^2 \angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1). \tag{45}$$

Proof. We use the proposed eigenexpansion of Eq. (42) to rewrite

$$\mathbf{B}\mathbf{w} = (\alpha_1 \mathbf{B}\mathbf{v}_1 + \mathbf{B}\mathbf{V}_2 \boldsymbol{\alpha}_2) \overset{(41)}{=} \alpha_1 \mathbf{B}\mathbf{v}_1 \overset{(38)}{=} \alpha_1 \lambda_1 \mathbf{S}\mathbf{v}_1 \tag{46}$$

and replace the above result into $\rho(\mathbf{w})$. Then

$$\begin{aligned} \rho(\mathbf{w}) = -\frac{\mathbf{w}^\top \mathbf{B}\mathbf{w}}{\mathbf{w}^\top \mathbf{S}\mathbf{w}} &\overset{(46)}{=} -\alpha_1 \lambda_1 \frac{(\alpha_1 \mathbf{v}_1 + \mathbf{V}_2 \boldsymbol{\alpha}_2)^\top \mathbf{S}\mathbf{v}_1}{\|\mathbf{w}\|^2_{\mathbf{S}}} \\ &\overset{(40)}{=} -\lambda_1 \frac{\alpha_1^2}{\|\mathbf{w}\|^2_{\mathbf{S}}} \overset{(43)}{=} -\lambda_1 \cos^2 \angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1), \end{aligned}$$

which proves the second part of the proposition. The first follows directly from property (P.1). □

Gradient-suboptimality connection Fermat's first-order optimality condition implies that the gradient is zero at the minimizer of $\rho(\mathbf{w})$. Considering the structure of $\rho(\mathbf{w})$, we propose a precise connection between the norm of gradient and suboptimality. Our analysis relies on the representation of the gradient $\nabla\rho(\mathbf{w})$ in the $\mathbf{S}$-orthonormal basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ which is described in the next proposition.

Proposition 2. *Using the* $\mathbf{S}$*-orthogonal basis as given in Eq.* (40)*, the gradient vector can be expanded as*

$$\begin{aligned}\|\mathbf{w}\|_{\mathbf{S}}^2 \nabla\rho(\mathbf{w})/2 = &- \lambda_1\alpha_1 \sin^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)\mathbf{S}\mathbf{v}_1 \\ &+ \lambda_1 \cos^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)\mathbf{S}\mathbf{V}_2\boldsymbol{\alpha}_2\end{aligned} \tag{47}$$

Proof. The above derivation is based on two results: (i) $\mathbf{v}_1$ is an eigenvector of $(\mathbf{B}, \mathbf{S})$ and (ii) the representation of $\rho(\mathbf{w})$ in Proposition 1. We recall the definition of $\nabla\rho(\mathbf{w})$ in (24) and write

$$\begin{aligned}\nabla\rho(\mathbf{w})\|\mathbf{w}\|_{\mathbf{S}}^2/2 &= -\rho(\mathbf{w})\mathbf{S}\mathbf{w} - \mathbf{B}\mathbf{w} \\ &\overset{(46)}{=} -\rho(\mathbf{w})\mathbf{S}\left(\alpha_1\mathbf{v}_1 + \mathbf{V}_2\boldsymbol{\alpha}_2\right) + \lambda_1\alpha_1\mathbf{S}\mathbf{v}_1 \\ &\overset{(45)}{=} -(1 - \cos^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1))\lambda_1\alpha_1\mathbf{S}\mathbf{v}_1 \\ &+\lambda_1 \cos^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)\mathbf{S}\mathbf{V}_2\boldsymbol{\alpha}_2 \\ &= -\lambda_1\alpha_1 \sin^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)\mathbf{S}\mathbf{v}_1 \\ &+\lambda_1 \cos^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)\mathbf{S}\mathbf{V}_2\boldsymbol{\alpha}_2\end{aligned} \tag{48}$$

□

Exploiting the gradient representation of the last proposition, the next proposition establishes the connection between suboptimality and the $\mathbf{S}^{-1}$-norm of gradient $\nabla_{\mathbf{w}}\rho(\mathbf{w})$.

Proposition 3. *Suppose that* $\rho(\mathbf{w}) \neq 0$*, then the* $\mathbf{S}^{-1}$*-norm of the gradient* $\nabla\rho(\mathbf{w})$ *relates to the suboptimality as*

$$\|\mathbf{w}\|_{\mathbf{S}}^2 \|\nabla\rho(\mathbf{w})\|_{\mathbf{S}^{-1}}^2/(4|\rho(\mathbf{w})|) = \rho(\mathbf{w}) - \rho(\mathbf{v}_1) \tag{49}$$

Proof. Multiplying the gradient representation in Proposition 2 by $\mathbf{S}^{-1}$ yields

$$\begin{aligned}\mathbf{S}^{-1}\nabla\rho(\mathbf{w})\|\mathbf{w}\|_{\mathbf{S}}^2/2 = &- \lambda_1\alpha_1 \sin^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)\mathbf{v}_1 \\ &+ \lambda_1 \cos^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)\mathbf{V}_2\boldsymbol{\alpha}_2.\end{aligned}$$

By combining the above result with the $\mathbf{S}$-orthogonality of the basis $(\mathbf{v}_1, \mathbf{V}_2)$, we derive the (squared) $\mathbf{S}^{-1}$-norm of the gradient as

$$\begin{aligned}&\nabla\rho(\mathbf{w})^\top\mathbf{S}^{-1}\nabla\rho(\mathbf{w})/4 \overset{(40)}{=} T_1 + T_2, \\ &T_1 := \|\mathbf{w}\|_{\mathbf{S}}^{-4}\lambda_1^2\alpha_1^2 \sin^4\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1), \\ &T_2 := \|\mathbf{w}\|_{\mathbf{S}}^{-4}\lambda_1^2 \cos^4\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)\|\boldsymbol{\alpha}_2\|^2.\end{aligned} \tag{50}$$

It remains to simplify the terms T_1 and T_2. For T_1,

$$\frac{\|\mathbf{w}\|_{\mathbf{S}}^2}{\lambda_1^2 \sin^4\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)} T_1 = \|\mathbf{w}\|_{\mathbf{S}}^{-2}\alpha_1^2 \overset{(43)}{=} \cos^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1).$$

Similarly, we simplify T_2:

$$\begin{aligned}\frac{\|\mathbf{w}\|_{\mathbf{S}}^2}{\lambda_1^2 \cos^4\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)} T_2 &= \|\mathbf{w}\|_{\mathbf{S}}^{-2}\|\boldsymbol{\alpha}_2\|^2 \\ &= \|\mathbf{w}\|_{\mathbf{S}}^{-2}\left(\|\mathbf{w}\|_{\mathbf{S}}^2 - \alpha_1^2\right) \\ &= 1 - \frac{\alpha_1^2}{\|\mathbf{w}\|_{\mathbf{S}}^2} \\ &\overset{(43)}{=} 1 - \cos^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1) \\ &= \sin^2\angle_{\mathbf{S}}(\mathbf{w}, \mathbf{v}_1)\end{aligned}$$

Replacing the simplified expression of T_1 and T_2 into Eq. (50) yields

$$\begin{aligned}\|\mathbf{w}\|_{\mathbf{S}}^2\|\nabla_{\mathbf{w}}\rho(\mathbf{w})\|_{\mathbf{S}^{-1}}^2/4 =&\lambda_1^2\cos^2\angle_{\mathbf{S}}(\mathbf{w},\mathbf{v}_1)\sin^2\angle_{\mathbf{S}}(\mathbf{w},\mathbf{v}_1)\left(\sin^2\angle_{\mathbf{S}}(\mathbf{w},\mathbf{v}_1)+\cos^2\angle_{\mathbf{S}}(\mathbf{w},\mathbf{v}_1)\right)\\ =&\lambda_1^2\cos^2\angle_{\mathbf{S}}(\mathbf{w},\mathbf{v}_1)\sin^2\angle_{\mathbf{S}}(\mathbf{w},\mathbf{v}_1),\\ \overset{(45)}{=}&|\rho(\mathbf{w})|\lambda_1\|\sin^2\angle_{\mathbf{S}}(\mathbf{w},\mathbf{v}_1)\\ \overset{(44)}{=}&|\rho(\mathbf{w})|\left(\rho(\mathbf{w})-\rho(\mathbf{v}_1)\right).\end{aligned}$$

A rearrangement of terms in the above equation concludes the proof. □

A.5 Convergence proof

We have seen: suboptimality in $\rho(\mathbf{w})$ directly relates to $\sin^2\angle_{\mathbf{S}}(\mathbf{w},\mathbf{v}_1)$ for all $\mathbf{w}\setminus\{\mathbf{0}\}$. In the next lemma we prove that this quantity is strictly decreased by repeated GD updates at a linear rate.

Lemma 3. *Suppose that Assumption 1 holds and consider* GD *(*GD*) steps on* (12) *with stepsize* $\eta_t=\frac{\|\mathbf{w}_t\|_{\mathbf{S}}^2}{2L|\rho(\mathbf{w}_t)|}$. *Then, for any* $\mathbf{w}_0\in\mathbb{R}^d$ *such that* $\rho(\mathbf{w}_0)<0$, *the updates of Eq.* (23) *yield the following linear convergence rate*

$$\sin^2\angle_{\mathbf{S}}(\mathbf{w}_t,\mathbf{v}_1)\leq\left(1-\frac{\mu}{L}\right)^{2t}\sin^2\angle_{\mathbf{S}}(\mathbf{w}_0,\mathbf{v}_1) \tag{51}$$

Proof. To prove the above statement, we relate the sine of the angle of a given iterate $\mathbf{w}_{t+1}$ with $\mathbf{v}_1$ in terms of the previous angle $\angle(\mathbf{w}_t,\mathbf{v}_1)$. Towards this end, we assume for the moment that $\rho(\mathbf{w}_t)\neq 0$ but note that this naturally always holds whenever $\rho(\mathbf{w}_0)\neq 0$,[6] such that the angle relation can be recursively applied through all $t\geq 0$ to yield Eq. (51).

(i) We start by deriving an expression for $\sin\angle_{\mathbf{S}}(\mathbf{v}_1,\mathbf{w}_t)$. By (30) and the definition $\rho(\mathbf{w}_t)$, we have that $-\rho(\mathbf{w}_t)\mathbf{w}_t$ is the $\mathbf{S}$-orthogonal projection of $\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t$ to $\operatorname{span}\{\mathbf{w}_t\}$. Indeed,

$$\begin{aligned}&\langle\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t+\rho(\mathbf{w}_t)\mathbf{w}_t,\mathbf{w}_t\rangle_{\mathbf{S}}\\ =&\mathbf{w}_t^\top\mathbf{B}\mathbf{S}^{-1}\mathbf{S}\mathbf{w}_t-\frac{\mathbf{w}_t^\top\mathbf{B}\mathbf{w}_t}{\mathbf{w}_t^\top\mathbf{S}\mathbf{w}_t}\mathbf{w}_t^\top\mathbf{S}\mathbf{w}_t=0.\end{aligned}$$

Note that $\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t=\left(\mathbf{S}^{-1}\mathbf{u}\right)\left(\mathbf{u}^\top\mathbf{w}_t\right)$ is a nonzero multiple of $\mathbf{v}_1$ and thus $\sin\angle_{\mathbf{S}}(\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t,\mathbf{w}_t)=\sin\angle_{\mathbf{S}}(\mathbf{v}_1,\mathbf{w}_t)$. Thus, by (P.2) we have

$$\begin{aligned}\sin\angle_{\mathbf{S}}(\mathbf{v}_1,\mathbf{w}_t)&=\sin\angle_{\mathbf{S}}(\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t,\mathbf{w}_t)\\ &=\frac{\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t+\rho(\mathbf{w}_t)\mathbf{w}_t\|_{\mathbf{S}}}{\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t\|_{\mathbf{S}}}.\end{aligned} \tag{52}$$

(ii) We now derive an expression for $\sin\angle_{\mathbf{S}}(\mathbf{v}_1,\mathbf{w}_{t+1})$. Let $a_{t+1}\in\mathbb{R}$ such that $a_{t+1}\mathbf{w}_{t+1}\in\mathbb{R}^d$ is the $\mathbf{S}$-orthogonal projection of $\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t$ to $\operatorname{span}\{\mathbf{w}_{t+1}\}$. Then

$$\begin{aligned}&\langle\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t-a_{t+1}\mathbf{w}_{t+1},\mathbf{w}_{t+1}\rangle_{\mathbf{S}}=0\\ \Rightarrow\ &\langle\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t-a_{t+1}\mathbf{w}_{t+1},(a_{t+1}+\rho(\mathbf{w}_t))\mathbf{w}_{t+1}\rangle_{\mathbf{S}}=0.\end{aligned} \tag{53}$$

By the Pythagorean theorem and (53), we get

$$\begin{aligned}\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t+\rho(\mathbf{w}_t)\mathbf{w}_{t+1}\|_{\mathbf{S}}^2&=\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t-a_{t+1}\mathbf{w}_{t+1}\|_{\mathbf{S}}^2\\ &\quad+\|(a_{t+1}+\rho(\mathbf{w}_t))\mathbf{w}_{t+1}\|_{\mathbf{S}}^2\\ &\geq\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t-a_{t+1}\mathbf{w}_{t+1}\|_{\mathbf{S}}^2.\end{aligned} \tag{54}$$

[6] as we will show later by induction.

Hence, again by (P.2)

$$\begin{aligned}\sin\angle_{\mathbf{S}}(\mathbf{v}_1,\mathbf{w}_{t+1}) &= \sin\angle_{\mathbf{S}}(\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t,\mathbf{w}_{t+1})\\ &= \frac{\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t - a_{t+1}\mathbf{w}_{t+1}\|_{\mathbf{S}}}{\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t\|_{\mathbf{S}}}\\ &\overset{(54)}{\leq} \frac{\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_{t+1}\|_{\mathbf{S}}}{\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t\|_{\mathbf{S}}}.\end{aligned} \tag{55}$$

(iii) To see how the two quantities on the right hand side of (52) and (55) relate, let us rewrite the GD updates from Eq. (23) as follows

$$\begin{aligned}&\rho(\mathbf{w}_t)\mathbf{w}_{t+1} = \rho(\mathbf{w}_t)\mathbf{w}_t - \frac{\mathbf{S}}{L}\left(\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_t\right)\\ \Leftrightarrow\ &\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_{t+1} = \mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_t - \frac{\mathbf{S}}{L}\left(\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_t\right)\\ \Leftrightarrow\ &\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_{t+1} = \left(\mathbf{I} - \frac{\mathbf{S}}{L}\right)\left(\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_t\right).\end{aligned} \tag{56}$$

By taking the $\mathbf{S}$-norm we can conclude

$$\begin{aligned}&\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_{t+1}\|_{\mathbf{S}}\\ \overset{(56)}{\leq}\ &\|\mathbf{I} - \frac{\mathbf{S}}{L}\|_{\mathbf{S}} \cdot \|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_t\|_{\mathbf{S}}\\ \leq\ &\left(1 - \frac{\mu}{L}\right)\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_t\|_{\mathbf{S}},\end{aligned} \tag{57}$$

where the first inequality is due to property (P.4) of the $\mathbf{S}$-spectral norm and the second is due to Assumption (1) and (P.3) , which allows us to bound the latter in term of the usual spectral norm as follows

$$\begin{aligned}\|\mathbf{I} - \mathbf{S}/L\|_{\mathbf{S}} &\overset{(P.3)}{=} \|\mathbf{S}^{1/2}\left(\mathbf{I} - \mathbf{S}/L\right)\mathbf{S}^{-1/2}\|_2\\ =\|\mathbf{I} - \mathbf{S}/L\|_2 &\overset{(2)}{\leq} 1 - \mu/L.\end{aligned} \tag{58}$$

(iv) Combining the above results yields the desired bound

$$\begin{aligned}\sin\angle_{\mathbf{S}}(\mathbf{v}_1,\mathbf{w}_{t+1}) &\overset{(55)}{\leq} \frac{\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_{t+1}\|_{\mathbf{S}}}{\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t\|_{\mathbf{S}}}\\ &\overset{(57)}{\leq} \left(1 - \frac{\mu}{L}\right)\frac{\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}_t)\mathbf{w}_t\|_{\mathbf{S}}}{\|\mathbf{S}^{-1}\mathbf{B}\mathbf{w}_t\|_{\mathbf{S}}}\\ &\overset{(52)}{=} \left(1 - \frac{\mu}{L}\right)\sin\angle_{\mathbf{S}}(\mathbf{v}_1,\mathbf{w}_t).\end{aligned} \tag{59}$$

(v) Finally, we show that the initially made assumption $\rho(\mathbf{w}_t) \neq 0$ is naturally satisfied in all iterations. First, $\rho(\mathbf{w}_0) < 0$ by assumption. Second, assuming $\rho(\mathbf{w}_{\hat{t}}) < 0$ for an arbitrary $\hat{t} \in \mathbb{N}^+$ gives that the above analysis (i-iv) holds for $\hat{t}+1$ and thus (59) together with (44) and the fact that $\lambda_1 > 0$ give

$$\begin{aligned}\rho(\mathbf{w}_{\hat{t}+1}) &= -\lambda_1(1 - \sin^2(\mathbf{v}_1,\mathbf{w}_{\hat{t}+1}))\\ &< -\lambda_1(1 - \sin^2(\mathbf{v}_1,\mathbf{w}_{\hat{t}})) = \rho(\mathbf{w}_{\hat{t}}) < 0,\end{aligned}$$

where the last inequality is our induction hypothesis. Thus $\rho(\mathbf{w}_{\hat{t}+1}) < 0$ and we can conclude by induction that $\rho(\mathbf{w}_t) < 0, \forall t \in \mathbb{N}^+$.

As a result, (59) holds $\forall t \in \mathbb{N}^+$, which (applied recursively) proves the statement (51). □

We are now ready to prove Theorem 1.

Proof of Theorem (1)*:* By combining the results of Lemma 3 as well as Proposition 1 and 3, we can complete the proof of the Theorem 1 as follows

$$\begin{aligned}
&\|\mathbf{w}_t\|_{\mathbf{S}}^2 \|\nabla_{\mathbf{w}} \rho(\mathbf{w}_t)\|_{\mathbf{S}^{-1}}^2 / |4\rho(\mathbf{w}_t)| \\
&\overset{(49)}{=} (\rho(\mathbf{w}_t) - \rho(\mathbf{w}^*)) \\
&\overset{(44)}{=} \lambda_1 \sin^2 \angle_{\mathbf{S}}(\mathbf{w}_t, \mathbf{v}_1) \\
&\overset{(51)}{\leq} \left(1 - \frac{\mu}{L}\right)^{2t} \lambda_1 \sin^2 \angle_{\mathbf{S}}(\mathbf{w}_0, \mathbf{v}_1) \\
&\overset{(44)}{=} \left(1 - \frac{\mu}{L}\right)^{2t} (\rho(\mathbf{w}_0) - \rho(\mathbf{w}^*)) .
\end{aligned}$$

B LEARNING HALFSPACES ANALYSIS

In this section, we provide a convergence analysis for Algorithm 1 on the problem of learning halfspaces

$$\min_{\tilde{\mathbf{w}} \in \mathbb{R}^d} \left(f_{\text{LH}}(\tilde{\mathbf{w}}) := \mathbf{E}_{y,\mathbf{x}} \left[\varphi(-y \mathbf{x}^\top \tilde{\mathbf{w}}) \right] = \mathbf{E}_{\mathbf{z}} \left[\varphi(\mathbf{z}^\top \tilde{\mathbf{w}}) \right] \right). \tag{16 revisited}$$

As before, we reparametrize $\tilde{\mathbf{w}}$ by means of the covariance matrix $\mathbf{S}$ as

$$\tilde{\mathbf{w}} := g \mathbf{w} / \|\mathbf{w}\|_{\mathbf{S}}. \tag{9 revisited}$$

We assume that the domain of $f_{\text{LH}}(\tilde{\mathbf{w}})$ is $\mathbb{R}^d$ but exclude $\mathbf{0}$ such that (9) is always well defined. Thus, the domain of the new parameterization is $(\mathbf{w}, g) \in \left(\mathbb{R}^d \setminus \{\mathbf{0}\}\right) \otimes \mathbb{R} \subset \mathbb{R}^{d+1}$.

B.1 Preliminaries

Recall the normality assumption on the data distribution.

Assumption 2. *[Normality assumption] We assume that* $\mathbf{z}$ *is a multivariate normal random variable distributed with mean* $\mathbf{E}[\mathbf{z}] = \mathbf{E}[-y\mathbf{x}] = \mathbf{u}$ *and second-moment* $\mathbf{E}\left[\mathbf{z}\mathbf{z}^\top\right] - \mathbf{E}[\mathbf{z}]\mathbf{E}[\mathbf{z}]^\top = \mathbf{E}\left[\mathbf{x}\mathbf{x}^\top\right] - \mathbf{u}\mathbf{u}^\top = \mathbf{S} - \mathbf{u}\mathbf{u}^\top$.

Under the above assumption we have that for a differentiable function $g(\mathbf{z}) : \mathbb{R}^d \to \mathbb{R}$ the following equality holds

$$\mathbf{E}_{\mathbf{z}}[g(\mathbf{z})\mathbf{z}] = \mathbf{E}_{\mathbf{z}}[g(\mathbf{z})]\,\mathbf{u} + (\mathbf{S} - \mathbf{u}\mathbf{u}^\top)\mathbf{E}_{\mathbf{z}}[\nabla_{\mathbf{z}} g(\mathbf{z})]. \tag{60}$$

This result, which can be derived using a simple application of integration by parts, is called *Stein's lemma* (Landsman and Nevslehová, 2008). In the next lemma, we show that this allows us to simplify the expression of the gradient of Eq. 16.

Lemma 4 (restated result from (Erdogdu et al., 2016))**.** *Under the normality assumption on the data distribution (Assumption 2), the gradient of* f_{LH} *(Eq. 16) can be expressed as*

$$\nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(\tilde{\mathbf{w}}) = c_1(\tilde{\mathbf{w}})\mathbf{u} + c_2(\tilde{\mathbf{w}})\mathbf{S}\tilde{\mathbf{w}}, \tag{61}$$

where $c_i \in \mathbb{R}$ *depends on the i-th derivative of the loss function denoted by* $\varphi^{(i)}$ *as*

$$\begin{aligned}
c_1(\tilde{\mathbf{w}}) &= \mathbf{E}_{\mathbf{z}}\left[\varphi^{(1)}(\mathbf{z}^\top \tilde{\mathbf{w}})\right] - \mathbf{E}_{\mathbf{z}}\left[\varphi^{(2)}(\mathbf{z}^\top \tilde{\mathbf{w}})\right](\mathbf{u}^\top \tilde{\mathbf{w}}), \\
c_2(\tilde{\mathbf{w}}) &= \mathbf{E}_{\mathbf{z}}\left[\varphi^{(2)}(\mathbf{z}^\top \tilde{\mathbf{w}})\right]
\end{aligned}$$

Proof. The gradient of f_{LH} can be written as follows

$$\nabla f_{LH} = \mathbf{E}\left[\varphi^{(1)}(\mathbf{z}^\top \tilde{\mathbf{w}})\mathbf{z}\right]. \tag{62}$$

A straight forward application of Stein's lemma (Eq. (60)) yields

$$\nabla f_{LH} = \mathbf{E}\left[\varphi^{(1)}(\mathbf{z}^\top \tilde{\mathbf{w}})\right]\mathbf{u} + (\mathbf{S} - \mathbf{u}\mathbf{u}^\top)\mathbf{E}\left[\varphi^{(2)}(\mathbf{z}^\top \tilde{\mathbf{w}})\right]\tilde{\mathbf{w}}, \tag{63}$$

which –after rearrangement – proves the result. See detailed derivation in (Erdogdu et al., 2016). □

In addition to the assumption on the data distribution, the proposed analysis also requires a rather weak assumption on f_{LH} and loss function φ.

Assumption 3. *[Assumptions on loss function] We assume that the loss function $\varphi : \mathbb{R} \to \mathbb{R}$ is infinitely differentiable, i.e. $\varphi \in C^\infty(\mathbb{R}, \mathbb{R})$, with a bounded derivative, i.e. $\exists \Phi > 0$ such that $|\varphi^{(1)}(\beta)| \leq \Phi, \forall \beta \in \mathbb{R}$.*

Assumption 4. *[Smoothness assumption] We assume that the objective $f : \mathbb{R}^d \to \mathbb{R}$ is ζ-smooth if it is differentiable on $\mathbb{R}$ and its gradient is ζ-Lipschitz. Furthermore, we assume that a solution $\alpha^* := \arg\min_\alpha \|\nabla f(\alpha \mathbf{w})\|^2$ exists that is bounded in the sense that $\forall \mathbf{w} \in \mathbb{R}^d, -\infty < \alpha^* < \infty$.*[7]

Recall that ζ-smoothness of f_{LH}, which is mentioned in the last assumption, implies that the gradient of f_{LH} is ζ-Lipschitz, i.e.

$$\|\nabla f_{\text{LH}}(\tilde{\mathbf{w}}_1) - \nabla f_{\text{LH}}(\tilde{\mathbf{w}}_2)\| \leq \zeta \|\tilde{\mathbf{w}}_1 - \tilde{\mathbf{w}}_2\|. \tag{64}$$

B.2 Global characterization

Here, we prove a result about a global property of the solution of the problem of learning halfspaces.

Lemma 1. *Under Assumptions 1 and 2, all bounded critical points $\tilde{\mathbf{w}}_*$ of f_{LH} have the general form*

$$\tilde{\mathbf{w}}_* = g_* \mathbf{S}^{-1}\mathbf{u},$$

where the scalar $g_* \in \mathbb{R}$ depends on $\tilde{\mathbf{w}}_*$ and the choice of the loss function φ.

Proof. Setting the gradient of the objective f_{LH} as given in Eq. (61) to zero directly gives the result. □

B.3 Established Convergence Rate

Based on this assumption, we derive a linear convergence rate for GDNP presented in Algorithm 1. We first restate the convergence guarantee before providing a detailed proof.

Theorem 4. *[Convergence rate of GDNP on learning halfspaces] Suppose Assumptions 1– 4 hold. Let $\tilde{\mathbf{w}}_{T_d}$ be the output of GDNP on f_{LH} with the following choice of stepsizes*

$$s_t := s(\mathbf{w}_t, g_t) = -\frac{\|\mathbf{w}_t\|_{\mathbf{S}}^3}{L g_t h(\mathbf{w}_t, g_t)} \tag{65}$$

for $t = 1, \ldots, T_d$, where

$$\begin{aligned} h(\mathbf{w}_t, g_t) :=& \mathbf{E}_{\mathbf{z}}\left[\varphi'\left(\mathbf{z}^\top \tilde{\mathbf{w}}_t\right)\right](\mathbf{u}^\top \mathbf{w}_t) \\ &- \mathbf{E}_{\mathbf{z}}\left[\varphi''\left(\mathbf{z}^\top \tilde{\mathbf{w}}_t\right)\right](\mathbf{u}^\top \mathbf{w}_t)^2 \end{aligned} \tag{66}$$

is a stopping criterion. If initialized such that $\rho(\mathbf{w}_0) \neq 0$ (see Eq. (12)), then $\tilde{\mathbf{w}}_{T_d}$ is an approximate critical point of f_{LH} in the sense that

$$\begin{aligned} \|\nabla_{\tilde{\mathbf{w}}} f(\tilde{\mathbf{w}}_{T_d})\|^2 \leq& (1 - \mu/L)^{2T_d} \Phi^2 \left(\rho(\mathbf{w}_0) - \rho^*\right) \\ &+ 2^{-T_s} \zeta |b_t^{(0)} - a_t^{(0)}| / \mu^2. \end{aligned} \tag{67}$$

[7] This is a rather technical but not so restrictive assumption. For example, it always holds for the sigmoid loss unless the classification error of $\mathbf{w}$ is already zero.

B.3.1 Proof sketch

As mentioned earlier, the objective f_{LH} on Gaussian inputs has a particular *global* property. Namely, all its critical points are aligned along the same direction. The key idea is that $\mathbf{S}$-reparameterization provides this global information to a local optimization method through an elegant length-direction decoupling. This allows GDNP to mimic the behaviour of Gradient Descent on the above mentioned Rayleigh quotient for the directional updates and thereby inherit the linear convergence rate. At the same time, the scaling factor can easily be brought to a critical point by a fast, one dimensional search algorithm. We formalize and combine these intuitions in a detailed proof below.

B.3.2 Gradient in the normalized parameterization

Since GDNP relies on the normalized parameterization, we first need to derive the gradient of the objective in this parameterization

$$\min_{\mathbf{w},g}\left(f_{\text{LH}}(\mathbf{w},g) := \mathbf{E}_{\mathbf{z}}\left[\varphi\left(g\frac{\mathbf{z}^\top\mathbf{w}}{\|\mathbf{w}\|_{\mathbf{S}}}\right)\right]\right). \tag{68}$$

Straight forward calculations yield the following connection between the gradient formulation in the original parameterization $\nabla_{\tilde{\mathbf{w}}} f_{\text{LH}} = \mathbf{E}_{\mathbf{z}}[\varphi^{(1)}(\tilde{\mathbf{w}}^\top\mathbf{z})\mathbf{z}]$ and the gradient in the normalized parameterization

$$\begin{aligned}\nabla_{\mathbf{w}} f_{\text{LH}}(\mathbf{w},g) &= g\mathbf{A}_{\mathbf{w}}\nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(\tilde{\mathbf{w}}),\\ \partial_g f_{\text{LH}}(\mathbf{w},g) &= \mathbf{w}^\top\nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(\tilde{\mathbf{w}})/\|\mathbf{w}\|_{\mathbf{S}}\end{aligned} \tag{69}$$

where

$$\mathbf{A}_{\mathbf{w}} := \mathbf{I}/\|\mathbf{w}\|_{\mathbf{S}} - \mathbf{S}\mathbf{w}\mathbf{w}^\top/\|\mathbf{w}\|_{\mathbf{S}}^3. \tag{70}$$

Note that the vector $\mathbf{S}\mathbf{w}$ is orthogonal to the column space of $\mathbf{A}_{\mathbf{w}}$ since

$$\mathbf{A}_{\mathbf{w}}\mathbf{S}\mathbf{w} = \left(\mathbf{S}\mathbf{w} - \frac{\|\mathbf{w}\|_{\mathbf{S}}^2}{\|\mathbf{w}\|_{\mathbf{S}}^2}\mathbf{S}\mathbf{w}\right)/\|\mathbf{w}\|_{\mathbf{S}} = 0. \tag{71}$$

We will repeatedly use the above property in our future analysis. In the next lemma, we establish a connection between the norm of gradients in different parameterizations.

Proposition 4. *Under the reparameterization* (9)*, the following holds:*

$$\begin{aligned}\|\nabla_{\tilde{\mathbf{w}}} f(\tilde{\mathbf{w}})\|_{\mathbf{S}^{-1}}^2 =& \|\mathbf{w}\|_{\mathbf{S}}^2\|\nabla_{\mathbf{w}} f(\mathbf{w},g)\|_{\mathbf{S}^{-1}}^2/g^2\\ &+ \left(\partial_g f(\mathbf{w},g)\right)^2\end{aligned} \tag{72}$$

Proof. We introduce the vector $\mathbf{q}_1 = \mathbf{S}\mathbf{w}/\|\mathbf{w}\|_{\mathbf{S}}$ that has unit $\mathbf{S}^{-1}$-norm, i.e. $\|\mathbf{q}_1\|_{S^{-1}} = 1$.

According to Eq. (71),

$$\mathbf{A}_{\mathbf{w}}\mathbf{q}_1 = 0$$

holds. Now, we extend this vector to an $\mathbf{S}^{-1}$-orthogonal basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_d\}$ of $\mathbb{R}^d$ such that

$$\langle\mathbf{q}_i, \mathbf{q}_i\rangle_{\mathbf{S}^{-1}} = 1,\ \forall i \text{ and } \langle\mathbf{q}_i, \mathbf{q}_j\rangle_{\mathbf{S}^{-1}} = 0,\ \forall i \neq j.$$

Let $\mathbf{Q}_2$ be a matrix whose columns are $\{\mathbf{q}_2, \ldots, \mathbf{q}_d\}$. The choice of $\mathbf{q}_1$ together with $\mathbf{S}^{-1}$-orthogonality of the basis imply that $\mathbf{w}$ is orthogonal to $\mathbf{Q}_2$:

$$\mathbf{w}^\top\mathbf{q}_j = \|\mathbf{w}\|_{\mathbf{S}}\langle\mathbf{q}_1, \mathbf{q}_j\rangle_{\mathbf{S}^{-1}} = 0, \forall j \neq 1$$

Consider the gradient expansion in the new basis, i.e.

$$\nabla_{\tilde{\mathbf{w}}} f(\tilde{\mathbf{w}}) = \alpha_1\mathbf{q}_1 + \mathbf{Q}_2\boldsymbol{\alpha}_2,\ \ \|\nabla_{\tilde{\mathbf{w}}} f(\tilde{\mathbf{w}})\|_{\mathbf{S}^{-1}}^2 = \alpha_1^2 + \|\boldsymbol{\alpha}_2\|_2^2$$

Plugging the above expansion into Eq. (69) yields

$$\nabla_{\mathbf{w}} f(\mathbf{w}, g) = g\mathbf{Q}_2\boldsymbol{\alpha}_2/\|\mathbf{w}\|_{\mathbf{S}}, \;\; \partial_g f(\mathbf{w}, g) = \alpha_1 \tag{73}$$

hence the $\mathbf{S}^{-1}$-norm of the directional gradient in the new parameterization is

$$\|\nabla_{\mathbf{w}} f(\mathbf{w}, g)\|^2_{\mathbf{S}^{-1}} = g^2\|\boldsymbol{\alpha}_2\|_2^2/\|\mathbf{w}\|^2_{\mathbf{S}} \tag{74}$$

Therefore, one can establish the following connection between the $\mathbf{S}^{-1}$-norm of gradient in the two different parameterizations:

$$\begin{aligned}\|\nabla_{\tilde{\mathbf{w}}} f(\tilde{\mathbf{w}})\|^2_{\mathbf{S}^{-1}} &= \alpha_1^2 + \|\boldsymbol{\alpha}_2\|_2^2 \\ &\overset{(73)}{=} (\partial_g f(\mathbf{w}, g))^2 + \|\boldsymbol{\alpha}_2\|^2 \\ &\overset{(74)}{=} (\partial_g f(\mathbf{w}, g))^2 + \frac{\|\mathbf{w}\|^2_{\mathbf{S}}\|\nabla_{\mathbf{w}} f(\mathbf{w}, g)\|^2_{\mathbf{S}^{-1}}}{g^2}\end{aligned}$$

□

The above lemma allows us to first analyze convergence in the $(\mathbf{w}, g)$-parameterization and then we relate the result to the original $\tilde{\mathbf{w}}$-parameterization in the following way: Given that the iterates $\{\mathbf{w}_t, g_t\}_{t\in\mathbf{N}^+}$ converge to a critical point of $f_{\text{LH}}(\mathbf{w}, g)$, one can use Eq. (73) to prove that $\tilde{\mathbf{w}}_t = g_t\mathbf{w}_t/\|\mathbf{w}_t\|_{\mathbf{S}}$ also converges to a critical point of $f_{\text{LH}}(\tilde{\mathbf{w}})$.

For the particular case of learning halfspaces with Gaussian input, the result of Lemma 4 allows us to write the gradient ∇f_{LH} as

$$\nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(\tilde{\mathbf{w}}) = c_1(\tilde{\mathbf{w}})\mathbf{u} + c_2(\tilde{\mathbf{w}})\mathbf{S}\tilde{\mathbf{w}}, \tag{75}$$

where the constants c_1 and c_2 are determined by the choice of the loss. Replacing this expression in Eq. (69) yields the following formulation for the gradient in normalized coordinates

$$\nabla_{\mathbf{w}} f_{\text{LH}}(\mathbf{w}) = gc_1(\mathbf{w}, g)\mathbf{A}_{\mathbf{w}}\mathbf{u} + gc_2(\mathbf{w}, g)\mathbf{A}_{\mathbf{w}}\mathbf{S}\mathbf{w}, \tag{76}$$

where $c_i(\mathbf{w}, g) = c_i(\tilde{\mathbf{w}}(\mathbf{w}, g))$. Yet, due to the specific matrix $\mathbf{A}_{\mathbf{w}}$ that arises when reparametrizing according to (9), the vector $\mathbf{S}\mathbf{w}$ is again in the kernel of $\mathbf{A}_{\mathbf{w}}$ (see Eq. (71)) and hence

$$\nabla_{\mathbf{w}} f_{\text{LH}}(\mathbf{w}) = gc_1(\mathbf{w}, g)\mathbf{A}_{\mathbf{w}}\mathbf{u}. \tag{77}$$

B.3.3 Convergence of the scalar g

Lemma 5 (Convergence of scalar)**.** *Under the assumptions of Theorem 4, in each iteration $t \in \mathbb{N}^+$ of* GDNP *(Algorithm 1) the partial derivative of f_{LH} as given in Eq.* (68) *converges to zero at the following linear rate*

$$\left(\partial_g f_{\text{LH}}(\mathbf{w}_t, a_t^{(T_s)})\right)^2 \leq 2^{-T_s}\zeta|b_t^{(0)} - a_t^{(0)}|/\mu^2. \tag{78}$$

Proof. According to Algorithm 1, the length of the search space for g is cut in half by each bisection step and thus reduces to

$$|a_t^{(T_\text{s})} - b_t^{(T_\text{s})}| \leq 2^{-T_\text{s}}|b_t^{(0)} - a_t^{(0)}|$$

after T_s iterations. The continuity of $\partial_g f_{\text{LH}}$ given by Assumption 4 and the fact that Algorithm 1 guarantees $\partial_g f_{\text{LH}}(\mathbf{w}_t, a_t^{(m)}) \cdot \partial_g f_{\text{LH}}(\mathbf{w}_t, b_t^{(m)}) < 0$, $\forall m \in \mathbb{N}^+$ allow us to conclude that there exists a root g^* for $\partial_g f_{\text{LH}}$ between $a_t^{(T_\text{s})}$ and $b_t^{(T_\text{s})}$ for which

$$|a_t^{(T_\text{s})} - g^*| \leq 2^{-T_\text{s}}|b_t^{(0)} - a_t^{(0)}|$$

holds.

The next step is to relate the above distance to the partial derivative of $f_{\text{LH}}(\mathbf{w}, g)$ w.r.t g. Consider the compact notation $\mathbf{w}'_t = \mathbf{w}_t / \|\mathbf{w}_t\|_{\mathbf{S}}$. Using this notation and the gradient expression in Eq. (69), the difference of partial derivatives can be written as

$$\begin{aligned}
&\left(\partial_g f_{\text{LH}}(\mathbf{w}_t, a_t^{(T_s)})\right)^2 \\
=&\left(\partial_g f_{\text{LH}}(\mathbf{w}_t, a_t^{(T_s)}) - \partial_g f_{\text{LH}}(\mathbf{w}_t, g^*)\right)^2 \\
=&\left(\left(\nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(a_t^{(T_s)}\mathbf{w}'_t) - \nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(g^*\mathbf{w}'_t)\right)^\top \mathbf{w}'_t\right)^2.
\end{aligned} \tag{79}$$

Using the smoothness assumption on f_{LH} we bound the above difference as follows

$$\begin{aligned}
&\left(\left(\nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(a_t^{(T_s)}\mathbf{w}'_t) - \nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(g^*\mathbf{w}'_t)\right)^\top \mathbf{w}'_t\right)^2 \\
\leq&\|\mathbf{w}'_t\|^2 \|\nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(a_t^{(T_s)}\mathbf{w}'_t) - \nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(g^*\mathbf{w}'_t)\|^2 \\
\overset{(64)}{\leq}& \zeta \|\mathbf{w}'_t\|^2 \|a_t^{(T_s)}\mathbf{w}'_t - g^*\mathbf{w}'_t\|^2 \\
\leq&\zeta \|\mathbf{w}'_t\|^4 (a_t^{(T_s)} - g^*)^2 \\
\leq&\zeta \|\mathbf{w}_t\|^4 \|\mathbf{w}_t\|_{\mathbf{S}}^{-4} (a_t^{(T_s)} - g^*)^2 \\
\overset{(2)}{\leq}&\zeta \mu^{-2} (a_t^{(T_s)} - g^*)^2,
\end{aligned} \tag{80}$$

where the last inequality is due to Assumption 1.

Combining (79) and (80) directly yields

$$\left(\partial_g f_{\text{LH}}(\mathbf{w}_t, a_t^{(T_s)})\right)^2 \leq 2^{-T_s} \zeta |b_t^{(0)} - a_t^{(0)}| / \mu^2, \tag{81}$$

which proves the assertion. □

B.3.4 Directional convergence

Lemma 6 (Directional convergence)**.** *Let all assumptions of Theorem 4 hold. Then, in each iteration $t \in \mathbb{N}^+$ of* GDNP *(Algorithm 1) with the following choice of stepsizes*

$$s_t := s(\mathbf{w}_t, g_t) = -\frac{\|\mathbf{w}_t\|_{\mathbf{S}}^3}{L g_t h(\mathbf{w}_t, g_t)}, \quad t = 1, \ldots, T_d \tag{82}$$

where

$$h(\mathbf{w}_t, g_t) := \mathbf{E}_{\mathbf{z}}\left[\varphi'\left(\tilde{\mathbf{w}}_t\right)\right]\left(\mathbf{u}^\top \mathbf{w}_t\right) - \mathbf{E}_{\mathbf{z}}\left[\varphi''(\tilde{\mathbf{w}}_t)\right]\left(\mathbf{u}^\top \mathbf{w}_t\right)^2 \neq 0. \tag{83}$$

The norm of the gradient w.r.t. $\mathbf{w}$ *of* f_{LH} *as in Eq.* (68) *converges at the following linear rate*

$$\|\mathbf{w}_t\|_{\mathbf{S}}^2 \|\nabla_{\mathbf{w}} f_{\text{LH}}(\mathbf{w}_t, g_t)\|_{\mathbf{S}^{-1}}^2 \leq (1 - \frac{\mu}{L})^{2t} \Phi^2 g_t^2 \left(\rho(\mathbf{w}_0) - \rho^*\right).$$

Proof. The key insight for this proof is a rather subtle connection between the gradient of the reparametrized

least squares objective (Eq. (12)) and the directional gradient of the learning halfspace problem (Eq. (68)):

$$\begin{aligned}
\nabla_{\mathbf{w}} f_{\text{LH}}(\mathbf{w}, g) &\overset{(77)}{=} g c_1(\mathbf{w}, g) \mathbf{A}_{\mathbf{w}} \mathbf{u} \\
&= g c_1(\mathbf{w}, g) \left(\mathbf{u} - \left(\frac{\mathbf{w}^\top \mathbf{u}}{\|\mathbf{w}\|_{\mathbf{S}}^2} \right) \mathbf{S}\mathbf{w} \right) / \|\mathbf{w}\|_{\mathbf{S}} \\
&= g \frac{c_1(\mathbf{w}, g)}{\mathbf{u}^\top \mathbf{w}} \left(\mathbf{u}^\top \mathbf{w} \mathbf{u} - \frac{(\mathbf{w}^\top \mathbf{u})^2}{\|\mathbf{w}\|_{\mathbf{S}}^2} \mathbf{S}\mathbf{w} \right) / \|\mathbf{w}\|_{\mathbf{S}} \\
&= g \frac{c_1(\mathbf{w}, g) \|\mathbf{w}\|_{\mathbf{S}}}{\mathbf{u}^\top \mathbf{w}} \left(\mathbf{B}\mathbf{w}_t + \rho(\mathbf{w}) \mathbf{S}\mathbf{w} \right) / \|\mathbf{w}\|_{\mathbf{S}}^2 \\
&\overset{(24)}{=} -g \left(\frac{c_1(\mathbf{w}, g) \|\mathbf{w}\|_{\mathbf{S}}}{2\mathbf{u}^\top \mathbf{w}} \right) \nabla_{\mathbf{w}} \rho(\mathbf{w}).
\end{aligned} \tag{84}$$

Therefore, the directional gradients $\nabla_{\mathbf{w}} \rho(\mathbf{w})$ and $\nabla_{\mathbf{w}} f_{\text{LH}}(\mathbf{w}, g)$ align in the same direction for all $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$. Based on this observation, we propose a stepsize schedule for GDNP such that we can exploit the convergence result established for least squares in Theorem 1. The iterates $\{\mathbf{w}_t\}_{t \in \mathbf{N}^+}$ of GDNP on f_{LH} can be written as

$$\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t - s_t \nabla_{\mathbf{w}} f_{\text{LH}}(\mathbf{w}_t, g) \\
&\overset{(84)}{=} \mathbf{w}_t + s_t \left(\frac{g_t c_1(\mathbf{w}_t, g_t) \|\mathbf{w}_t\|_{\mathbf{S}}}{2\mathbf{u}^\top \mathbf{w}_t} \right) \nabla_{\mathbf{w}} \rho(\mathbf{w}_t).
\end{aligned} \tag{85}$$

The stepsize choice of Eq.(82) guarantees that

$$s_t \left(\frac{g_t c_1(\mathbf{w}_t, g_t) \|\mathbf{w}_t\|_{\mathbf{S}}}{2\mathbf{u}^\top \mathbf{w}_t} \right) = -\frac{\|\mathbf{w}_t\|_{\mathbf{S}}^4}{(2L(\mathbf{w}_t^\top \mathbf{u})^2)} \overset{(25)}{=} -\eta_t.$$

Thus, Eq. (85) can be rewritten as

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla_{\mathbf{w}} \rho(\mathbf{w}_t),$$

which exactly matches the GD iterate sequence of Eq. (13) on $\rho(\mathbf{w})$. At this point, we can invoke the result of Theorem 1 to establish the following convergence rate:

$$\begin{aligned}
&\|\mathbf{w}_t\|_{\mathbf{S}}^2 \|\nabla_{\mathbf{w}} f_{\text{LH}}(\mathbf{w}_t, g_t)\|_{\mathbf{S}^{-1}}^2 \\
\overset{(84)}{=} &\|\mathbf{w}_t\|_{\mathbf{S}}^2 c_1^2(\mathbf{w}_t, g_t) g_t^2 \|\nabla_{\mathbf{w}} \rho(\mathbf{w}_t)\|_{\mathbf{S}^{-1}}^2 / \left(2(\mathbf{u}^\top \mathbf{w}_t)/\|\mathbf{w}_t\|_{\mathbf{S}}\right)^2 \\
\overset{(3)}{\leq} &\Phi^2 \|\mathbf{w}_t\|_{\mathbf{S}}^2 g_t^2 \|\nabla_{\mathbf{w}} \rho(\mathbf{w}_t)\|_{\mathbf{S}^{-1}}^2 / \left(2(\mathbf{u}^\top \mathbf{w}_t)/\|\mathbf{w}_t\|_{\mathbf{S}}\right)^2 \\
\leq &\Phi^2 \|\mathbf{w}_t\|_{\mathbf{S}}^2 g_t^2 \|\nabla_{\mathbf{w}} \rho(\mathbf{w}_t)\|_{\mathbf{S}^{-1}}^2 / |4\rho(\mathbf{w}_t)| \\
\overset{(15)}{\leq} &\Phi^2 g_t^2 \left(\rho(\mathbf{w}_t) - \rho^*\right) \\
\overset{(14)}{\leq} &(1 - \mu/L)^{2t} \Phi^2 g_t^2 \left(\rho(\mathbf{w}_0) - \rho^*\right).
\end{aligned} \tag{86}$$

□

B.3.5 Combined convergence guarantee

Using Proposition 4 and combining the results obtained for optimizing the directional and scalar components, we finally obtain the following convergence guarantee:

$$\begin{aligned}
\|\nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(\tilde{\mathbf{w}}_{T_{\text{d}}})\|_{\mathbf{S}^{-1}}^2 &\overset{(72)}{=} \|\mathbf{w}_{T_{\text{d}}}\|_{\mathbf{S}}^2 \|\nabla_{\mathbf{w}} f_{\text{LH}}(\mathbf{w}_{T_{\text{d}}}, g_{T_{\text{d}}})\|_{\mathbf{S}^{-1}}^2 / g_{T_{\text{d}}}^2 \\
&\quad + \left(\partial_g f_{\text{LH}}(\mathbf{w}_{T_{\text{d}}}, g_{T_{\text{d}}})\right)^2 \\
&\overset{(86)}{\leq} (1 - \mu/L)^{2T_{\text{d}}} \Phi^2 \left(\rho(\mathbf{w}_0) - \rho^*\right) \\
&\quad + \left(\partial_g f_{\text{LH}}(\mathbf{w}_{T_{\text{d}}}, g_{T_{\text{d}}})\right)^2 \\
&\overset{(81)}{\leq} (1 - \mu/L)^{2T_{\text{d}}} \Phi^2 \left(\rho(\mathbf{w}_0) - \rho^*\right) \\
&\quad + 2^{-T_{\text{s}}} \zeta |b_{T_{\text{d}}}^{(0)} - a_{T_{\text{d}}}^{(0)}| / \mu^2.
\end{aligned}$$

Algorithm 3 Bisection

Input: T_{s}, $a_t^{(0)}$, $b_t^{(0)}$, f
Choose $a_t^{(0)}$ and $b_t^{(0)}$ such that $\partial_g f(a_t^{(0)}, \mathbf{w}_t) \cdot \partial_g f(b_t^{(0)}, \mathbf{w}_t) > 0$.
for $m = 0, \dots, T_{\text{s}}$ **do**
  $c = (a^{(m)} + b^{(m)})/2$
  if $\partial_g f(c, \mathbf{w}_t) \cdot \partial_g f(a^{(m)}, \mathbf{w}_t) > 0$ **then**
    $a^{(m+1)} \leftarrow c$
  else
    $b^{(m+1)} \leftarrow c$
  end if
end for
$g \leftarrow a^{(T_{\text{s}})}$
return g

B.3.6 A word on Weight Normalization

The improved convergence rate for Batch Normalization (Theorem 4) relies heavily on the fact that normalizing and backpropagating through the variance term resembles splitting the optimization task into a length- and directional component. As mentioned in the introduction, this feature is also present in Weight Normalization and it is thus an obvious question, whether Wn can achieve a similar convergence rate. From a theoretical perspective, we were not able to prove this which is essentially due to the subtle difference in how the normalization is done: While Bn normalizes the parameters to live on the **S**-sphere, Wn brings all parameters to the unit sphere.

The fast directional convergence rate of Bn on Learning Halfspaces is essentially inherited from the fast convergence of Gradient Descent (with adaptive stepsize) on the Rayleight Quotient. This can be seen in the proof of Lemma 6 where we specifically use the fact that $\nabla_{\mathbf{w}} f_{LH}$ and $\nabla_{\mathbf{w}} \rho$ align in the same direction. To prove this fact we need two ingredients (i) Stein's Lemma which gives us the expression of $\nabla_{\tilde{\mathbf{w}}} f_{LH}$ as in Eq. (75) and (ii) the specific reparametrization of Bn as in Eq. (9) which lets us express the directional part of this gradient as $\nabla_{\tilde{\mathbf{w}}} f_{LH} = g \mathbf{A}_{\mathbf{w}} \nabla_{\tilde{\mathbf{w}}} f_{\text{LH}}(\tilde{\mathbf{w}})$. As we shall see, the second part is very specific to the reparametrization done by Bn, which gives certain properties of $\mathbf{A}_{\mathbf{w}} = \mathbf{I}/\|\mathbf{w}\|_{\mathbf{S}} - \mathbf{S}\mathbf{w}\mathbf{w}^\top/\|\mathbf{w}\|_{\mathbf{S}}^3$ that then yield Eq. (77) which is simply a scaled version of the Rayleigh Quotient gradient (see Eq. (24)). This fact arises particularly because (i) $\mathbf{A}_{\mathbf{w}}$ is orthogonal to $\mathbf{S}\mathbf{w}$ (see Eq. (70)) and (ii) $\mathbf{A}_{\mathbf{w}}$ and $\nabla_{\mathbf{w}} \rho$ both involve division by the **S**-norm. Both properties are not given for the version of $\mathbf{A}_{\mathbf{w},\text{WN}} = \mathbf{I}/\|\mathbf{w}\|_2 - \mathbf{w}\mathbf{w}^\top/\|\mathbf{w}\|_2^3$ that would arise when using Weight Normalization so the proof strategy breaks because we no longer match the gradients $\nabla_{\mathbf{w}} f_{LH}$ and $\nabla_{\mathbf{w}} \rho$.

That said, we observe similar empirical convergence behaviour in terms of suboptimality for Bn and Wn (without any adaptive stepsizes, see Section 4.4) but as can be seen on the right of Figure 7 the path that the two methods take can be very different. We thus leave it as an interesting open question if other settings and proof strategies can be found where fast rates for Wn are provable.

C NEURAL NETWORKS

Recall the training objective of the one layer MLP presented in Section 5:

$$\min_{\tilde{\mathbf{W}}, \Theta} \left(f_{\text{NN}}(\tilde{\mathbf{W}}, \Theta) := \mathbf{E}_{y,\mathbf{x}} \left[\ell \left(-yF(\mathbf{x}, \tilde{\mathbf{W}}, \Theta) \right) \right] \right), \tag{Revisited 20}$$

where

$$F(\mathbf{z}, \tilde{\mathbf{W}}, \Theta)) := \sum_{i=1}^{m} \theta^{(i)} \varphi(\mathbf{z}^\top \tilde{\mathbf{w}}^{(i)}).$$

Figure 4 illustrates the considered architecture in this paper.

Since the activation function is assumed to be an odd function (tanh), this choice allows us to equivalently rewrite the training objective as

$$\min_{\tilde{\mathbf{W}}, \Theta} \left(f_{\text{NN}}(\tilde{\mathbf{W}}, \Theta) = \mathbf{E}_{\mathbf{z}} \left[\ell(F \left(\mathbf{z}, \tilde{\mathbf{W}}, \Theta) \right) \right] \right). \tag{87}$$

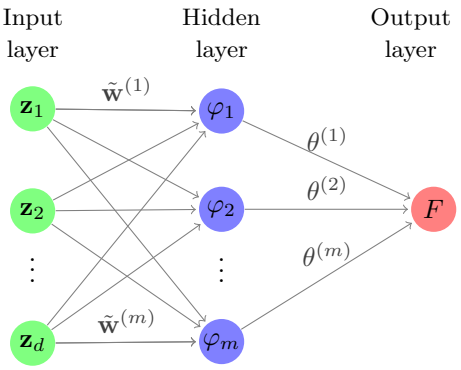

Figure 4: Neural network architecture considered in this paper.

By means of Assumption 2 and Stein's lemma (Eq. (60)) we can simplify the gradient w.r.t $\tilde{\mathbf{w}}_i$ as follows

$$\nabla_{\tilde{\mathbf{w}}^{(i)}} f_{\text{NN}}(\tilde{\mathbf{W}}, \Theta)/\theta^{(i)} = \mathbf{E}_{\mathbf{z}}\left[\ell^{(1)}(F(\mathbf{z}, \tilde{\mathbf{W}}, \Theta))\varphi^{(1)}(\mathbf{z}^\top \tilde{\mathbf{w}}^{(i)})\mathbf{z}\right] \tag{88}$$

$$= \alpha^{(i)}\mathbf{u} + \beta^{(i)}\mathbf{S}\tilde{\mathbf{w}}^{(i)} + \sum_{j=1}^{m} \gamma^{(i,j)}\mathbf{S}\tilde{\mathbf{w}}^{(j)}, \tag{89}$$

where the scalars $\alpha^{(i)}$, $\beta^{(i)}$ and $\gamma^{(i,j)}$ are defined as

$$\beta^{(i)} := \mathbf{E}_{\mathbf{z}}\left[\ell^{(1)}(F(\mathbf{z}, \tilde{\mathbf{W}}, \Theta))\varphi^{(2)}(\mathbf{z}^\top \tilde{\mathbf{w}}^{(i)})\right] \tag{90}$$

$$\gamma^{(i,j)} := \theta^{(j)}\mathbf{E}_{\mathbf{z}}\left[\ell^{(2)}(F(\mathbf{z}, \tilde{\mathbf{W}}, \Theta))\varphi^{(1)}(\mathbf{z}^\top \tilde{\mathbf{w}}^{(i)})\varphi^{(1)}(\mathbf{z}^\top \tilde{\mathbf{w}}^{(j)})\right] \tag{91}$$

$$\alpha^{(i)} := \mathbf{E}_{\mathbf{z}}\left[\ell^{(1)}(F(\mathbf{z}, \tilde{\mathbf{W}}, \Theta))\varphi^{(1)}(\mathbf{z}^\top \tilde{\mathbf{w}}^{(i)})\right] - \sum_{j=1}^{m} \gamma^{(i,j)}(\mathbf{u}^\top \tilde{\mathbf{w}}^{(j)}), \tag{92}$$

where $l^{(i)}(\cdot) \in \mathbb{R}$ and $\varphi^{(i)}(\cdot) \in \mathbb{R}$ represent the i-th derivative of $l(\cdot)$ and $\varphi(\cdot)$ with respect to their input $(\cdot)$.

C.1 Characterization of the objective

Interestingly, the normality assumption induces a particular global property on $f_{\text{NN}}(\tilde{\mathbf{W}})$. In fact, all critical weights $\tilde{\mathbf{w}}_i$ align along one single line in $\mathbb{R}^d$, which only depends on *incoming* information into the hidden layer. This result is formalized in the next lemma.

Lemma 2. *Suppose Assumptions 1 and 2 hold and let $\hat{\mathbf{w}}^{(i)}$ be a critical point of $f_{\text{NN}}(\tilde{\mathbf{W}})$ with respect to hidden unit i and for a fixed $\Theta \neq \mathbf{0}$. Then, there exits a scalar $\hat{c}^{(i)} \in \mathbb{R}$ such that*

$$\hat{\mathbf{w}}^{(i)} = \hat{c}^{(i)}\mathbf{S}^{-1}\mathbf{u}, \quad \forall i = 1, \ldots, m. \tag{21}$$

Proof. Recall the gradient of f_{NN} as given in Eq. (88). Computing a first order critical point requires setting the derivatives of all units to zero which amounts to solving the following system of non-linear equations:

$$\begin{aligned}
&(1) \quad \alpha^{(1)}\mathbf{u} + \beta^{(1)}\mathbf{S}\hat{\mathbf{w}}^{(1)} + \sum_{j=1}^{m} \gamma^{(1,j)}\mathbf{S}\hat{\mathbf{w}}^{(j)} &&= 0 \\
&(2) \quad \alpha^{(2)}\mathbf{u} + \beta^{(2)}\mathbf{S}\hat{\mathbf{w}}^{(2)} + \sum_{j=1}^{m} \gamma^{(2,j)}\mathbf{S}\hat{\mathbf{w}}^{(j)} &&= 0 \\
&\qquad\qquad \vdots \\
&(m) \quad \alpha^{(m)}\mathbf{u} + \beta^{(m)}\mathbf{S}\hat{\mathbf{w}}^{(m)} + \sum_{j=1}^{m} \gamma^{(m,j)}\mathbf{S}\hat{\mathbf{w}}^{(j)} &&= 0,
\end{aligned} \tag{93}$$

where each row (i) represents a system of d equations.

Matrix formulation of system of equations Let us rewrite (93) in matrix form. Towards this end, we define

$$\begin{aligned}\mathbf{U} &= [\mathbf{u}, \mathbf{u}, \dots, \mathbf{u}] \in \mathbb{R}^{d\times m},\\ \hat{\mathbf{w}} &= [\hat{\mathbf{w}}^{(1)}, \hat{\mathbf{w}}^{(2)}, \dots, \hat{\mathbf{w}}^{(m)}] \in \mathbb{R}^{d\times m}\end{aligned}$$

as well as

$$\begin{aligned}\mathbf{A} &= \operatorname{diag}\left(\alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(m)}\right) \in \mathbb{R}^{m\times m},\\ \mathbf{B} &= \operatorname{diag}\left(\beta^{(1)}, \beta^{(2)}, \dots, \beta^{(m)}\right) \in \mathbb{R}^{m\times m}\end{aligned}$$

and

$$\mathbf{\Gamma} = \begin{bmatrix} \gamma^{(1,1)} & \gamma^{(1,2)} & \dots & \gamma^{(1,m)} \\ \gamma^{(2,1)} & \gamma^{(2,2)} & \dots & \gamma^{(2,m)} \\ & & \ddots & \\ \gamma^{(m,1)} & \gamma^{(m,2)} & \dots & \gamma^{(m,m)}. \end{bmatrix}.$$

Note that $\mathbf{\Gamma} = \mathbf{\Gamma}^\top$ since $\gamma^{(i,j)} = \gamma^{(j,i)}, \forall i, j$.

Solving the system of equations Using the notation introduced above, we can write (93) as follows

$$\begin{aligned}&\mathbf{U}\mathbf{A} + \mathbf{S}\hat{\mathbf{W}}\mathbf{B} + \mathbf{S}\hat{\mathbf{W}}\mathbf{\Gamma} = 0\\ \Leftrightarrow &\mathbf{S}\hat{\mathbf{W}}(\mathbf{B}+\mathbf{\Gamma}) = -\mathbf{U}\mathbf{A}\\ \Leftrightarrow &\hat{\mathbf{W}} = -\mathbf{S}^{-1}\mathbf{U}\underbrace{\mathbf{A}(\mathbf{B}+\mathbf{\Gamma})^\dagger}_{:=\mathbf{D}},\end{aligned} \tag{94}$$

where $(\mathbf{B}+\mathbf{\Gamma})^\dagger$ is the pseudo-inverse of $(\mathbf{B}+\mathbf{\Gamma})$.

As a result

$$\begin{aligned}&[\hat{\mathbf{w}}^{(1)}, \hat{\mathbf{w}}^{(2)}, \dots, \hat{\mathbf{w}}^{(m)}]\\ =&-[\mathbf{S}^{-1}\mathbf{u}, \mathbf{S}^{-1}\mathbf{u}, \dots, \mathbf{S}^{-1}\mathbf{u}][\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_m]\end{aligned}$$

and hence the critical points of the objective are of the following type

$$\hat{\mathbf{w}}^{(i)} = -\mathbf{S}^{-1}\mathbf{U}\mathbf{d}_1 = -\left(\sum_{k=1}^{m}\mathbf{d}_i^{(k)}\right)\mathbf{S}^{-1}\mathbf{u}. \tag{95}$$

□

C.2 Possible implications for deep neural networks

From Eq. (21) in the Lemma 2 we can conclude that the optimal direction of any $\hat{\mathbf{w}}^{(i)}$ is independent of the corresponding output weight $\theta^{(i)}$, which only affects $\hat{\mathbf{w}}^{(i)}$ through the scaling parameter $\hat{c}^{(i)}$. This is a very appealing property: Take a multilayer network and assume (for the moment) that all layer inputs are Gaussian. Then, Lemma 2 still holds for any given hidden layer and gives rise to a decoupling of the optimal direction of this layer with all downstream weights, which in turn simplifies the curvature structure of the network since many Hessian blocks become zero.

However, classical local optimizers such as GD optimize both, direction and scaling, at the same time and are therefore blind to the above global property. It is thus very natural that performing optimization in the reparametrized weight space can in fact benefit from splitting the subtasks of optimizing scaling and direction in two parts, since updates in the latter are no longer sensitive to changes in the downstream part of the network. In the next section, we theoretically prove that such a decoupling accelerates optimization of weights of each individual unit in the presence of Gaussian inputs. Of course, the normality assumption is very strong but remarkably the experimental results of Section 5.3 suggest the validity of this result beyond the Gaussian design setting and thus motivate future research in this direction.

C.3 Convergence analysis

Here, we prove the convergence result restated below.

Theorem 3. *[Convergence of* GDNP *on MLP] Suppose Assumptions 1– 4 hold. We consider optimizing the weights* $(\mathbf{w}^{(i)}, g^{(i)})$ *of unit* i*, assuming that all directions* $\{\mathbf{w}^{(j)}\}_{j<i}$ *are critical points of* f_{NN} *and* $\mathbf{w}^k = \mathbf{0}$ *for* $k > i$*. Then,* GDNP *with step-size policy* $s^{(i)}$ *as in* (100) *and stopping criterion* $h^{(i)}$ *as in* (101) *yields a linear convergence rate on* $f^{(i)}$ *in the sense that*

$$\begin{aligned}\|\nabla_{\tilde{\mathbf{w}}^{(i)}} f(\tilde{\mathbf{w}}_t^{(i)})\|_{\mathbf{S}^{-1}}^2 \leq & (1-\mu/L)^{2t} C\left(\rho(\mathbf{w}_0) - \rho^*\right) \\ & + 2^{-T_s^{(i)}} \zeta |b_t^{(0)} - a_t^{(0)}|/\mu^2,\end{aligned} \tag{22}$$

where the constant $C > 0$ *is defined in Eq.* (104)*.*

Proof. According to the result of Lemma 2, all critical points of f_{NN} are aligned along the same direction as the solution of normalized least-squares. This property is similar to the objective of learning halfspaces (with Gaussian inputs) and the proof technique below therefore follows similar steps to the convergence proof of Theorem 4.

Gradient in the original parameterization Recall the gradient of f_{NN} is defined as

$$\nabla_{\tilde{\mathbf{w}}^{(i)}} f^{(i)}/\theta^{(i)} = \alpha^{(i)}\mathbf{u} + \beta^{(i)}\mathbf{S}\tilde{\mathbf{w}}^{(i)} + \sum_{j=1}^{m} \gamma^{(i,j)}\mathbf{S}\tilde{\mathbf{w}}^{(j)}. \tag{88 revisited}$$

Gradient in the normalized parameterization: Let us now consider the gradient of f_{NN} w.r.t the normalized weights, which relates to the gradient in the original parameterization in the following way

$$\begin{aligned}\nabla_{\mathbf{w}} f(\mathbf{w}, g) &= g\mathbf{A}_{\mathbf{w}} \nabla_{\tilde{\mathbf{w}}} f(\tilde{\mathbf{w}}), \\ \partial_g f(\mathbf{w}, g) &= \frac{\mathbf{w}^\top \nabla_{\tilde{\mathbf{w}}} f(\tilde{\mathbf{w}})}{\|\mathbf{w}\|_{\mathbf{S}}}\end{aligned} \tag{69 revisited}$$

Replacing the expression given in Eq. (88) into the above formula yields

$$\begin{aligned}\nabla_{\mathbf{w}^{(i)}} f_{\text{NN}}/(g^{(i)}\theta^{(i)}) = & \alpha^{(i)}\mathbf{A}_{\mathbf{w}^{(i)}}\mathbf{u} + \beta^{(i)}\mathbf{A}_{\mathbf{w}^{(i)}}\mathbf{S}\mathbf{w}^{(i)} \\ & + \sum_{j=1}^{m} \gamma^{(i,j)}\mathbf{A}_{\mathbf{w}^{(i)}}\mathbf{S}\mathbf{w}^{(j)},\end{aligned} \tag{96}$$

where

$$\mathbf{A}_{\mathbf{w}^{(i)}} := \mathbf{I}/\|\mathbf{w}^{(i)}\|_{\mathbf{S}} - \mathbf{S}\mathbf{w}^{(i)} \otimes \mathbf{w}^{(i)}/\|\mathbf{w}^{(i)}\|_{\mathbf{S}}^3. \tag{97}$$

Note that the constants $\alpha^{(i)}$, $\beta^{(i)}$ and $\gamma^{(i,j)}$ all depend on the parameters $\mathbf{w}^{(i)}$ and $\theta^{(j)}$ of the respective units i and j. The orthogonality of $\mathbf{S}\mathbf{w}^{(i)}$ to $\mathbf{A}_{\mathbf{w}^{(i)}}$ (see Eq. (71)) allows us to simplify things further:

$$\nabla_{\mathbf{w}^{(i)}} f_{\text{NN}}/(g^{(i)}\theta^{(i)}) = \alpha^{(i)}\mathbf{A}_{\mathbf{w}^{(i)}}\mathbf{u} + \sum_{j\neq i} \gamma^{(i,j)}\mathbf{A}_{\mathbf{w}^{(i)}}\mathbf{S}\mathbf{w}^{(j)} \tag{98}$$

We now use the initialization of weights $\{\mathbf{w}^{(k)} = c_k\mathbf{S}^{-1}\mathbf{u}\}_{k<i}$ and $\{\mathbf{w}^{(j)} = \mathbf{0}\}_{j>i}$ into the above expression to get

$$\begin{aligned}\nabla_{\mathbf{w}^{(i)}} f_{\text{NN}} &= \theta^{(i)} g^{(i)} \xi_t \mathbf{A}_{\mathbf{w}^{(i)}}\mathbf{u}, \quad \xi = \alpha^{(i)} + \sum_{j<i} \gamma^{(i,j)} c_j \\ &= \theta^{(i)} g^{(i)} \xi \left(\|\mathbf{w}^{(i)}\|_{\mathbf{S}}/(2\mathbf{u}^\top \mathbf{w}^{(i)})\right) \nabla\rho(\mathbf{w}^{(i)})\end{aligned} \tag{99}$$

where $\nabla\rho(\mathbf{w})$ is the gradient of the normalized ordinary least squares problem (Eq. (12)), i.e.

$$-\nabla\rho(\mathbf{w})/2 = \left(\mathbf{u}\mathbf{u}^\top\mathbf{w} + \frac{(\mathbf{u}^\top\mathbf{w})^2}{\|\mathbf{w}\|_{\mathbf{S}}^2}\mathbf{S}\mathbf{w}\right)/\|\mathbf{w}\|_{\mathbf{S}}^2.$$

We conclude that the global characterization property described in Eq. (21) transfers to the gradient since the above gradient aligns with the gradient of $\rho(\mathbf{w})$.

Choice of stepsize and stopping criterion We follow the same approach used in the proof for learning halfspaces and choose a stepsize to ensure that the gradient steps on $f_{\text{NN}}^{(i)}$ match the gradient iterates on ρ, i.e.

$$\begin{aligned}\mathbf{w}_{t+1}^{(i)} &= \mathbf{w}_t^{(i)} - s_t^{(i)}\theta^{(i)}g_t^{(i)}\xi_t\left(\|\mathbf{w}_t^{(i)}\|_{\mathbf{S}}/(2\mathbf{u}^\top\mathbf{w}^{(i)})\right)\nabla\rho(\mathbf{w}_t^{(i)}) \\ &= \mathbf{w}_t^{(i)} - \frac{\|\mathbf{w}_t\|_{\mathbf{S}}^2}{2L|\rho(\mathbf{w}_t^{(i)})|}\nabla\rho(\mathbf{w}_t^{(i)}),\end{aligned}$$

which leads to the following choice of stepsize

$$\begin{aligned}s_t^{(i)} &= \|\mathbf{w}_t^{(i)}\|_{\mathbf{S}}^3/(L\theta^{(i)}g_t^{(i)}\xi_t\mathbf{u}^\top\mathbf{w}_t^{(i)}) \\ \xi_t &= \alpha_t^{(i)} + \sum_{j<i}\gamma_t^{(i,j)}c_j.\end{aligned} \tag{100}$$

If $\xi_t = 0$, then the gradient is zero. Therefore, we choose the stopping criterion as follows

$$h_t^{(i)} = \xi_t = \alpha_t^{(i)} + \sum_{j<i}\gamma_t^{(i,j)}c_j. \tag{101}$$

Gradient norm decomposition Proposition 4 relates the $\mathbf{S}^{-1}$-norm of the gradient in the original space to the normalized space as follows

$$\begin{aligned}\|\nabla_{\tilde{\mathbf{w}}^{(i)}}f(\tilde{\mathbf{w}}_t^{(i)})\|_{\mathbf{S}^{-1}}^2 = \|\mathbf{w}_t^{(i)}\|_{\mathbf{S}}^2\|\nabla_{\mathbf{w}^{(i)}}f(\mathbf{w}_t^{(i)}, g^{(i)})\|_{\mathbf{S}^{-1}}^2/(g_t^{(i)})^2 \\ + \left(\partial_{g^{(i)}}f(\mathbf{w}_t^{(i)}, g_t^{(i)})\right)^2.\end{aligned} \tag{72 revisited}$$

In the following, we will establish convergence individually in terms of g and $\mathbf{w}$ and then use the above result to get a global result.

Convergence in scalar $g^{(i)}$ Since the smoothness property defined in Assumption 4 also holds for $f_{\text{NN}}^{(i)}$, we can directly invoke the result of Lemma 5 to establish a convergence rate for g:

$$\left(\partial_{g^{(i)}}f^{(i)}(\mathbf{w}_t^{(i)}, g_{T_{\text{s}}}^{(i)})\right)^2 \leq 2^{-T_{\text{s}}^{(i)}}\zeta|b_t^{(0)} - a_t^{(0)}|/\mu^2 \tag{81 revisited}$$

Directional convergence By the choice of stepsize in Eq. (100), the gradient trajectory on f_{NN} reduces to the gradient trajectory on $\rho(\mathbf{w})$. Hence, we can establish a linear convergence in $\mathbf{w}^{(i)}$ by a simple modification of Eq. (86):

$$\|\mathbf{w}_t^{(i)}\|_{\mathbf{S}}^2\|\nabla_{\mathbf{w}}f^{(i)}(\mathbf{w}_t^{(i)}, g_t^{(i)})\|_{\mathbf{S}^{-1}}^2 \leq \quad (1-\mu/L)^{2t}\xi_t^2 g_t^2\left(\rho(\mathbf{w}_0) - \rho(\mathbf{w}^*)\right). \tag{102}$$

The assumption 3 on loss with the choice of activation function as tanh allows us to bound the scalar ξ_t^2:

$$\xi_t^2 \leq 2\Phi^2 + 2i\sum_{j<i}(\theta^{(j)}c_j)^2. \tag{103}$$

Combined convergence bound Combining the above results concludes the proof in the following way

$$\begin{aligned}\|\nabla_{\tilde{\mathbf{w}}^{(i)}}f(\tilde{\mathbf{w}}_t^{(i)})\|_{\mathbf{S}^{-1}}^2 \leq &(1-\mu/L)^{2t}C\left(\rho(\mathbf{w}_0) - \rho^*\right) \\ &+ 2^{-T_{\text{s}}^{(i)}}\zeta|b_t^{(0)} - a_t^{(0)}|/\mu^2,\end{aligned}$$

where

$$C = 2\Phi^2 + 2i\sum_{j<i}(\theta^{(j)}c_j)^2 > 0. \tag{104}$$

□

D EXPERIMENTAL DETAILS

D.1 Learning Halfspaces

Setting We consider empirical risk minimization (ERM) as a surrogate for (16) in the binary classification setting and make two different choices for $\varphi(\cdot)$:

$$\textit{softplus}(\mathbf{w}^\top \mathbf{z}) := \mathbf{E}_{\mathbf{z}} \left[\log(1 + \exp(\tilde{\mathbf{w}}^\top \mathbf{z})) \right],$$
$$\textit{sigmoid}(\mathbf{w}^\top \mathbf{z}) := \mathbf{E}_{\mathbf{z}} \left[1/(1 + \exp(-\tilde{\mathbf{w}}^\top \mathbf{z})) \right]$$

The first resembles classical convex logistic regression when $\mathbf{y}_i \in \{-1, 1\}$. The second is a commonly used non-convex, continuous approximation of the zero-one loss in learning halfspaces (Zhang et al., 2015)

As datasets we use the common realworld dataset *a9a* ($n = 32'561, d = 123$) as well a synthetic data set drawn from a multivariate gaussian distribution such that $\mathbf{z} \sim \mathcal{N}(\mathbf{u}, \mathbf{S})$ ($n = 1'000, d = 50$).

Methods We compare the convergence behavior of GD and Accelerated Gradient Descent (AGD) (Nesterov, 2013) to Batch Normalization plus two versions of GD as well as Weight Normalization. Namely, we assess

- GDNP as stated in Algorithm 1 but with the Bisection search replaced by multiple Gradient Descent steps on g (10 per outer iteration)
- Batch Norm plus standard GD which simultaneously updates $\mathbf{w}$ and $\mathbf{g}$ with one gradient step on each parameter.
- Weight Normalization plus standard GD as above. (Salimans and Kingma, 2016)

All methods use full batch sizes. GDNP uses stepsizes according to the policy proposed in Theorem 4. On the softplus, GD, AGD, WN and BN are run with their own, constant, grid-searched stepsizes. Weight- and Batch Norm use a different stepsize for direction and scaling but only take one gradient step in each parameter per iteration. Since the sigmoid setting is non-convex and many different local minima and saddle points may be approched by the different algorithms in different runs, there exist no meaningful performance measure to gridsearch the stepsizes. We thus pick the inverse of the gradient Lipschitz constant ζ for all methods and all parameters, except GDNP. To estimate ζ, we compute $\zeta_{\text{sup}} := \|\mathbf{Z}^\top \mathbf{Z}\|_2 / 10 \geq \sum_i \varphi(\mathbf{w}^\top \mathbf{z}_i)^{(2)} \|\mathbf{Z}^\top \mathbf{Z}\|_2 / n$ where $\mathbf{Z} := [z_1, \ldots, z_n]^\top \in \mathbb{R}^{n \times d}$ and $\varphi(\cdot)^{(2)} \leq 0.1$. After comparison with the largest eigenvalue of the Hessian at a couple of thousand different parameters $\tilde{\mathbf{w}}$ we found the bound to be pretty tight. Note that for GDNP, we use $L := \|\mathbf{Z}^\top \mathbf{Z}\|_2$ which can easily be computed as a pre-processing step and is – contrary to ζ – independent of $\mathbf{w}$. AGD computes the momentum parameter $\beta_t = (t-2)/(t+1)$ in the convex case and uses a (grid-searched) constant $\beta_t = \beta \in [0, 1]$ in the non-convex setting. We initialize randomly, i.e. we draw $\tilde{\mathbf{w}}_0 \sim \mathcal{N}(0, 1)$, set $\mathbf{w}_0 := \tilde{\mathbf{w}}_0$ and choose g_0 sucht that $\tilde{\mathbf{w}}_0 = g \mathbf{w}_0 / \|\mathbf{w}_0\|_{\mathbf{S}}$.

Results The Gaussian design experiments clearly confirm Theorem 4 in the sense that the loss in the convex-, as well as the gradient norm in the non-convex case decrease at a linear rate. The results on *a9a* show that GDNP can accelerate optimization even when the normality assumption does not hold and in a setting where no covariate shift is present. This motivates future research of non-linear reparametrizations even in convex optimization.

Regarding BN and WN we found a clear trade-off between making fast progress in the beginning and optimizing the last couple of digits. In the above results of Figure 2 and 5 we report runs with stepsizes that were optimized for the latter case but we here note that early progress can easily be achieved in normalized parametrizations (which the linear a9a softplus plot actually confirms) e.g. by putting a higher learning rate on $\mathbf{g}$. In the long run similar performance to that of GD sets in, which suggests that the length-direction decoupling does not fully do the trick. The superior performance of GDNP points out that either an increased number of steps in the scaling factor g or an adaptive stepsize scheme such as the one given in Eq. (25) (or both) may significantly increase the performance of Batch Normalized Gradient Descent BN.

It is thus an exciting open question whether such simple modifications to GD can also speed up the training of Batch Normalized neural networks. Finally, since GDNP performs similar to AGD in the non-gaussian setting, it

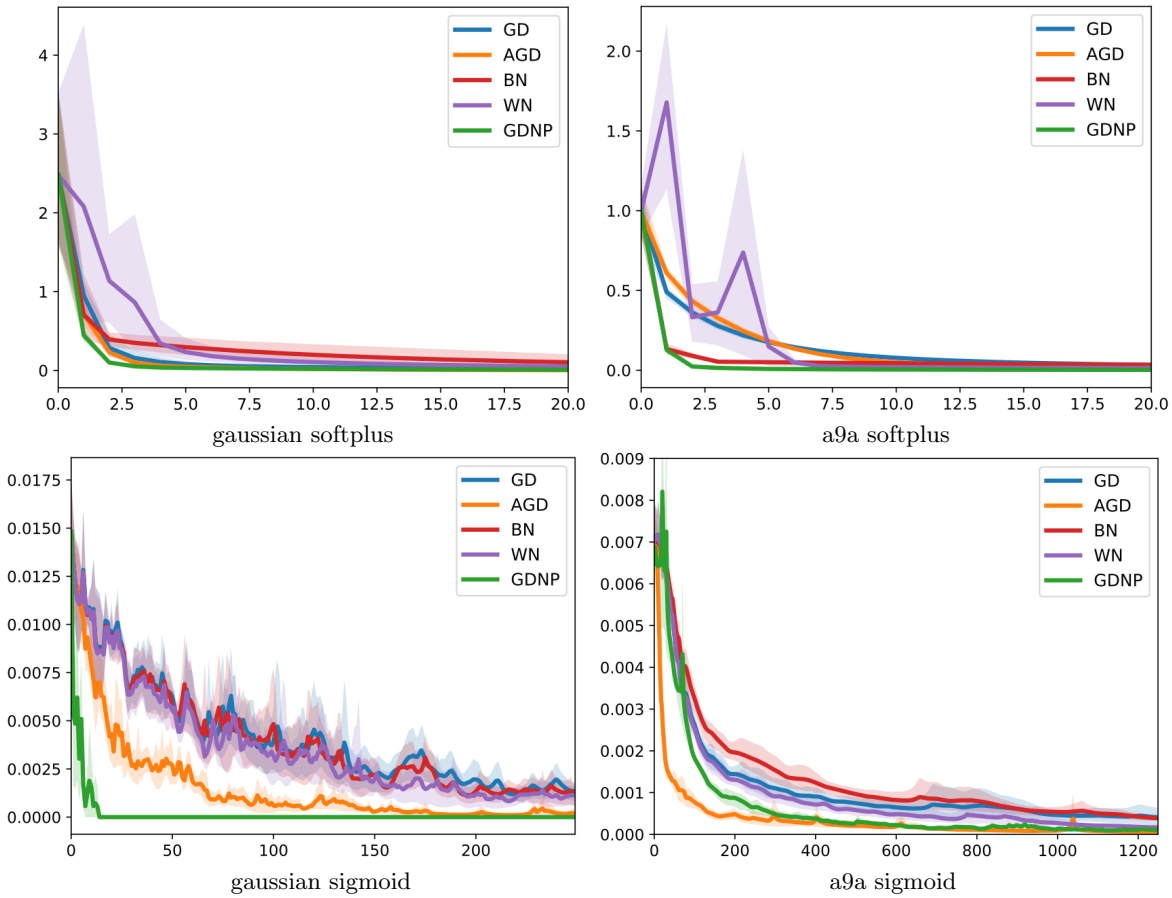

Figure 5: The plots are the same as in Figure 2 but show results in linear instead of log terms: Results of an average run (solid line) in terms of log suboptimality (softplus) and log gradient norm (sigmoid) over iterations as well as 90% confidence intervals of 10 runs with random initialization.

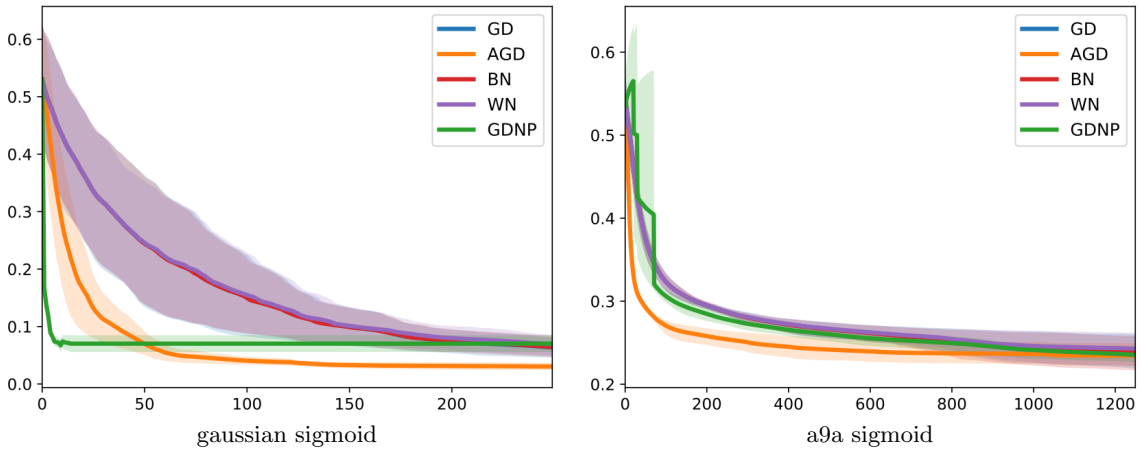

Figure 6: Addition to Figure 2 and 5: Suboptimality on the non-convex sigmoid problems in linear terms.

is a logical next step to study how accelerated gradient methods like AGD or Heavy Ball perform in normalized coordinates.

As a side note, Figure 7 shows how surprisingly different the paths that Gradient Descent takes before and after normalization can be.

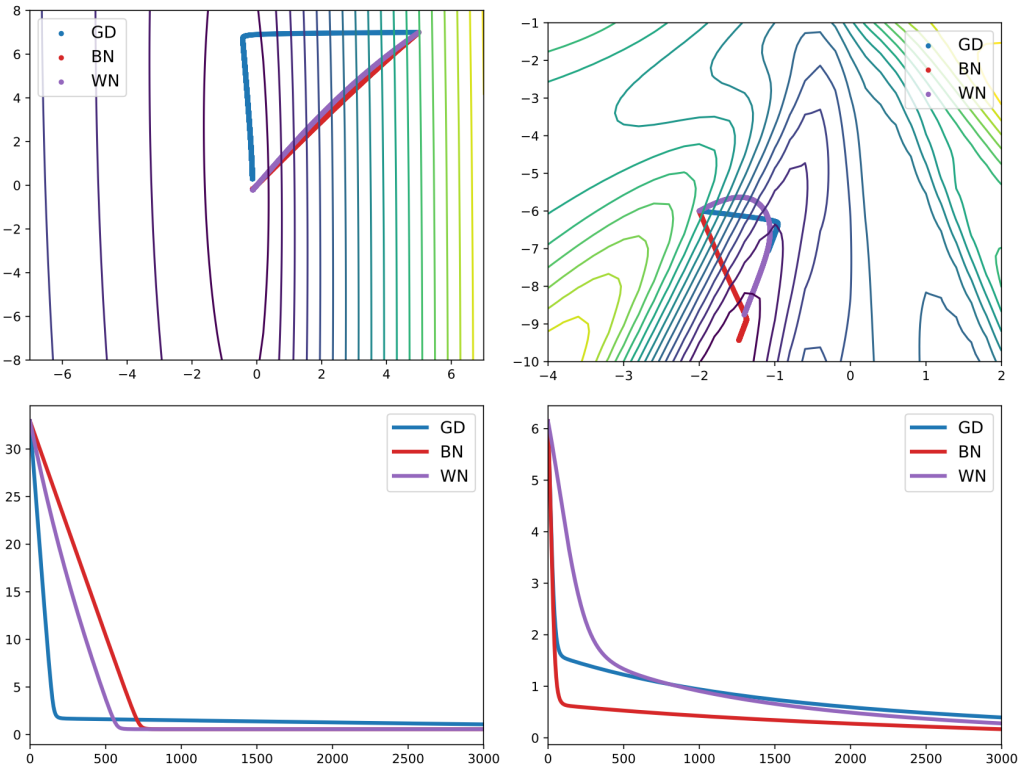

Figure 7: Normalization can lead to suprisingly different paths: Level sets and path (top) as well as sub-optimality (bottom) of GD, BN and WN (with constant step size and fixed number of iterations) on two instances of learning halfspaces with Gaussian data ($n = 5000$, $d = 2$). Left: convex logistic regression, right: non-convex sigmoidal regression.

D.2 Neural networks

Setting and methods We test the validity of Theorem 3 and Lemma 2 outside the Gaussian setting and a normalized and an unnormalized feedforward networks on the CIFAR10 image classification task. This dataset consists of 60000 32x32 images in 10 classes, with 6000 images per class (Krizhevsky and Hinton, 2009). The networks have six hidden layers with 50 hidden units in each of them. Each hidden unit has a *tanh* activation function, except for the very last layer which is linear. These scores are fed into a cross entropy loss layer which combines softmax and negative log likelihood loss. The experiments are implemented using the PyTorch framework (Paszke et al., 2017).

The first network is trained by standard GD and the second by GD in normalized coordinates (i.e. BN) with the same fixed stepsize on and $\mathbf{w}$, but we increase the learning rate on $\mathbf{g}$ by a factor of 10 which accelerates training significantly. The second network thus resembles performing standard GD in a network where all hidden layers are Batch Normalized. We measure the cross-dependency of the central with all other layers in terms of the Frobenius norm of the second partial derivatives $\frac{\partial^2 f_{\text{NN}}}{\partial \mathbf{W}_4 \partial \mathbf{W}_i}$. This quantity signals how the gradients of layer 4 change when we alter the direction of any other layer. From an optimization perspective, this is a sound measure for the cross-dependencies: If it is close to zero (high), that means that a change in layer i induces no (a large) change in layer 4. Compared to gradient calculations, computing second derivatives is rather expensive $O(nd^2)$ (where $d = 66700$), which is why we evaluate this measure every only 250 iterations.

Results Figure 3 and 8 confirm that the directional gradients of the central layer are affected far more by the upstream than by the downstream layers to a surprisingly large extent. Interestingly, this holds even before reaching a critical point. The cross-dependencies are generally decaying for the Batch Normalized network (BN) while they remain elevated in the un-normalized network (GD), which suggest that using Batch Normalization layers indeed simplifies the networks curvature structure in $\mathbf{w}$ such that the length-direction decoupling allows Gradient Descent to exploit simpler trajectories in these normalized coordinates for faster convergence. Of course, we cannot untangle this effect fully from the covariate shift reduction that was mentioned in the introduction.

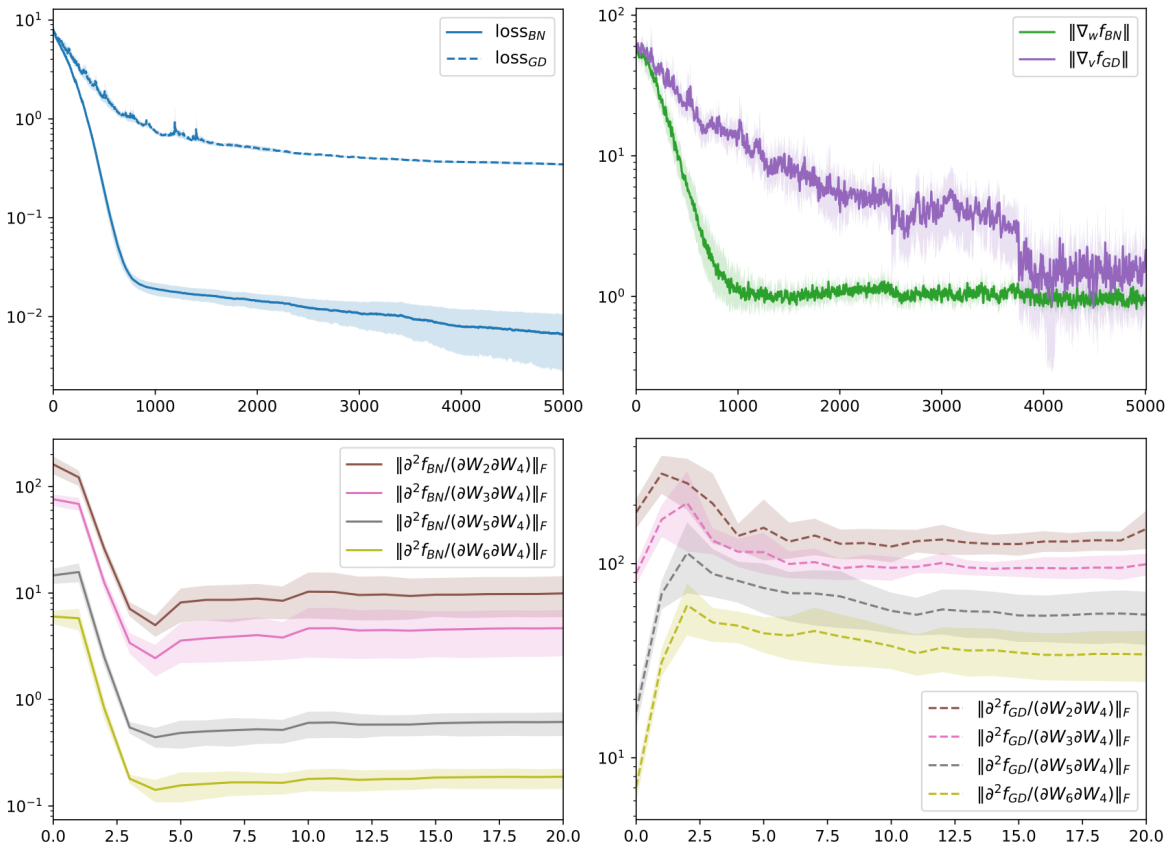

Figure 8: The plots are the same as in Figure 3 but show results in log instead of linear terms: (i) Loss, (ii) gradient norm and dependencies between central- and all other layers for BN (iii) and GD (iv) on a 6 hidden layer network with 50 units (each) on the CIFAR10 dataset.

Yet, the fact that the (de-)coupling increases in the distance to the middle layer (note how earlier (later) layers are more (less) important for the $\mathbf{W}_4$) emphasizes the relevance of this analysis particularly for deep neural network structures, where downstream dependencies might vanish completely with depth. This does not only make gradient based training easier but also suggests the possibility of using partial second order information, such as diagonal Hessian approximations (e.g. proposed in (Martens et al., 2012)).