

# Sparse Training with Lipschitz Continuous Loss Functions and a Weighted Group $l_0$ -norm Constraint

**Michael R. Metel**  
*Huawei Noah's Ark Lab*  
*Montréal, Qc, Canada*

MICHAEL.METEL@HUAWEI.COM

**Editor:** Silvia Villa

## Abstract

This paper is motivated by structured sparsity for deep neural network training. We study a weighted group  $l_0$ -norm constraint, and present the projection and normal cone of this set. Using randomized smoothing, we develop zeroth and first-order algorithms for minimizing a Lipschitz continuous function constrained by any closed set which can be projected onto. Non-asymptotic convergence guarantees are proven in expectation for the proposed algorithms for two related convergence criteria which can be considered as approximate stationary points. Two further methods are given using the proposed algorithms: one with non-asymptotic convergence guarantees in high probability, and the other with asymptotic guarantees to a stationary point almost surely. We believe in particular that these are the first such non-asymptotic convergence results for constrained Lipschitz continuous loss functions.

**Keywords:** structured sparsity,  $L_0$ -norm, Lipschitz continuity, stochastic optimization, non-asymptotic convergence

## 1. Introduction

This paper focuses on training deep neural networks with structured sparsity using a constrained optimization approach. A structured sparsity constraint allows for a simpler neural network architecture to be selected which can be deployed on low-resource devices. Research on sparsity in deep learning is vast, for a thorough background see (Hoeffler et al., 2021). Though much of this research is of a heuristic nature, our focus is on algorithms with theoretical convergence guarantees. The problem is modelled as

$$\min_{w \in \mathbb{R}^d} f(w) \quad \text{s.t. } w \in C \quad (1)$$

where  $C$  is a weighted group  $l_0$ -norm constraint defined in Section 3. This work examines the Euclidean projection operator and the normal cone of  $C$ , and develops new non-asymptotic convergence results for general zeroth and first-order stochastic projected algorithms for assumptions on  $f(w)$  applicable for a wide range of architectures in deep learning (Davis et al., 2020). In particular, the function  $f$  is only assumed to be Lipschitz continuous on a compact set, taking the form of the expected value of an integrable stochastic loss function  $F(w, \xi)$ ,  $f(w) := \mathbb{E}[F(w, \xi)]$ , where  $\xi \in \mathbb{R}^p$  is a random vector from a probability space  $(\Omega, \mathcal{F}, P)$ . In the context of supervised learning, given samples  $\xi^i = (x^i, y^i)$  for  $i = 1, 2, \dots, \mathcal{M}$ , where  $\{x^i\}$  is a feature set,  $\{y^i\}$  is a label set, and  $F(w, \xi^i)$  is the loss associated with sample  $i$ ,

$f(w)$  can be replaced in (1) by its approximation  $\hat{f}(w) = \mathcal{M}^{-1} \sum_{i=1}^{\mathcal{M}} F(w, \xi^i)$ .

The next section summarizes the required definitions and notation which will be used throughout the paper. Section 3 presents the weighted group  $l_0$ -norm constraint. Section 4 gives an overview of related works focusing on algorithms with theoretical convergence guarantees for Lipschitz continuous loss functions. Section 5 gives the detailed assumptions on  $F(w, \xi)$ , and presents the technique of using randomized smoothing to overcome the non-differentiability of the loss function. In Section 6 the Euclidean projection operator and the normal cone for the proposed constraint set are given. Section 7 presents the Stochastic Projected Algorithm (SPA), which has a zeroth and a first-order version, with new non-asymptotic convergence results for two related convergence criteria, and a method using SPA which has an asymptotic convergence guarantee to a stationary point almost surely. Section 8 shows how backpropagation can be used in conjunction with the first-order version of SPA for a wide range of deep learning architectures and validates its use in Section 9 where the theory of Section 7 is applied to train a neural network. Section 10 concludes the work. All proofs of results can be found in the Appendices A-E.

## 2. Preliminaries

For a set  $S$ , let the notation  $x \xrightarrow{S} w$  mean  $x \rightarrow w$  with  $x \in S$ , and for a discontinuous function  $h$ ,  $x \xrightarrow{h} w$  indicates that  $x \rightarrow w$  with  $h(x) \rightarrow h(w)$ . For a function  $h : \mathbb{R}^d \rightarrow \mathbb{R}$ ,

$$\limsup_{x \rightarrow w} h(x) := \inf_{\gamma > 0} \left( \sup_{0 < \|x-w\|_2 < \gamma} h(x) \right).$$

For a set-valued mapping  $G : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ ,

$$\text{Lim sup}_{x \rightarrow w} G(x) := \{y \in \mathbb{R}^d : \exists \text{ sequences } x_k \rightarrow w \text{ and } y_k \rightarrow y \text{ with } y_k \in G(x_k) \forall k \in \mathbb{N}\}.$$

For  $w \in S$ , the Fréchet normal cone equals

$$\widehat{N}(w, S) := \{y \in \mathbb{R}^n : \limsup_{x \xrightarrow{S} w} \frac{\langle y, x - w \rangle}{\|x - w\|_2} \leq 0\}$$

and the Mordukhovich normal cone equals

$$N(w, S) = \text{Lim sup}_{x \xrightarrow{S} w} \widehat{N}(x, S).$$

When  $w \notin S$ ,  $N(w, S) = \widehat{N}(w, S) := \{\emptyset\}$ . For more information about normal cones see for example Mordukhovich (2013).

For an extended real-valued function  $h : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\pm\infty\}$ , when finite, let  $\widehat{\partial}h(w)$  denote its Fréchet subdifferential, defined as

$$\widehat{\partial}h(w) := \{y \in \mathbb{R}^n : \liminf_{x \rightarrow w} \frac{f(x) - f(w) - \langle y, x - w \rangle}{\|x - w\|_2} \geq 0\},$$

and let  $\partial h(w)$  denote its Mordukhovich subdifferential,

$$\partial h(w) := \text{Lim sup}_{x \xrightarrow{h} w} \widehat{\partial} h(x).$$

Assuming that  $f(w)$  is locally Lipschitz continuous and  $S$  is closed, a necessary condition for  $\bar{w}$  to be locally optimal for the problem

$$\min_{w \in \mathbb{R}^d} f(w) \quad \text{s.t. } w \in S$$

is for (Rockafellar and Wets, 2009, Theorem 8.15 & 9.13)

$$0 \in \partial f(\bar{w}) + N(\bar{w}, S). \tag{2}$$

As a non-asymptotic convergence criterion for optimization algorithms, for an  $\epsilon > 0$ ,  $\bar{w}$  is an  $\epsilon$ -stationary point when

$$\text{dist}(0, \partial f(\bar{w}) + N(\bar{w}, S)) \leq \epsilon.$$

Let  $\bar{\partial} f(w)$  denote the Clarke subdifferential which equals  $\bar{\partial} f(w) = \text{co}\{\partial f(w)\}$ , where  $\text{co}\{\cdot\}$  is the convex hull, given that  $f(w)$  is locally Lipschitz continuous (Rockafellar and Wets, 2009, Theorem 9.61). When  $f(w)$  is Clarke regular, meaning that its one-sided directional derivative exists and for all  $v \in \mathbb{R}^d$   $f'(w; v) = \max_{g \in \bar{\partial} f(w)} \langle g, v \rangle$  (Clarke, 1990, Proposition 2.1.2 (b) & Definition 2.3.4), the Clarke subdifferential coincides with the Fréchet and Mordukhovich subdifferentials (Rockafellar and Wets, 2009, Theorem 9.61 & Corollary 8.11).

In this work, we will consider a relaxed version of (2), which we call a Clarke-Mordukhovich (C-M) stationary point:

$$0 \in \bar{\partial} f(\bar{w}) + N(\bar{w}, S). \tag{3}$$

Let  $B(x, r) := \{x + z : \|z\|_2 < r\}$  be the open Euclidean ball centered at  $x$  with radius  $r$ , let  $\bar{B}(x, r)$  be the corresponding closed Euclidean ball, and let  $\bar{B}_r := \{z \in \mathbb{R}^d : |z_i| \leq r \text{ for } i = 1, 2, \dots, d\}$  denote the closed  $l_\infty$ -ball with radius  $r > 0$  centered at 0.

We will also consider the Clarke  $\epsilon$ -subdifferential,

$$\bar{\partial}_\epsilon f(w) := \text{co}\{\bar{\partial} f(x) : x \in \bar{B}(w, \epsilon)\},$$

which was introduced in (Goldstein, 1977). This type of subdifferential has recently been used in the non-asymptotic convergence analysis of minimization algorithms for unconstrained Lipschitz continuous functions, see (Zhang et al., 2020; Metel and Takeda, 2022; Kornowski and Shamir, 2021) for more background. Besides its use for non-asymptotic convergence analysis, it holds that  $\lim_{\epsilon \rightarrow 0} \bar{\partial}_\epsilon f(\bar{w}) = \bar{\partial} f(\bar{w})$  (Zhang et al., 2020, Lemma 7), which motivates the proposed C-M stationary point (3) for our asymptotic convergence analysis.

The indicator function of a set  $S$  equals

$$\delta_S(w) = \begin{cases} 0 & \text{if } w \in S \\ \infty & \text{otherwise,} \end{cases}$$

and  $2^S$  denotes its power set. For a random variable  $X$ , let  $P_X$  denote the probability measure induced by the random variable  $X$ , i.e. for a Borel set  $S$ ,  $P_X(S) = P(\{\omega \in \Omega : X(\omega) \in S\})$ . For an  $n \in \mathbb{N}$ , let  $[n] := \{1, 2, \dots, n\}$  and  $[n]_{-1} := \{0, 1, \dots, n-1\}$ . When studying the computational complexity of algorithms we will use the notation  $\tilde{O}$  which is the standard big O notation with logarithmic terms ignored, e.g.  $\log^k(x) = \tilde{O}(1)$  for any  $k \in \mathbb{R}$ .

### 3. Weighted group $l_0$ -norm constraint

The  $l_0$ -norm counts the number of non-zero elements in a vector  $w \in \mathbb{R}^d$ ,

$$\|w\|_0 := \sum_{i=1}^d \mathbb{1}_{\{w \in \mathbb{R}^d : w_i \neq 0\}}(w).$$

For an  $n \leq d$ , let  $\{w^i\}_{i=1}^n$  be a partition of  $w$ , where  $w^i$  is of dimension  $d_i$  for each  $i \in [n]$  and  $\sum_{i=1}^n d_i = d$ . The weighted group  $l_0$ -norm constraint is then defined as

$$C := \{w \in \mathbb{R}^d : \sum_{i=1}^n p_i \mathbb{1}_{\{w^i \neq 0\}}(w) \leq m\},$$

where  $p_i > 0$  is a finite penalty associated with the subset of decision variables  $w^i$ ,  $\{w^i \neq 0\}$  denotes the set  $\{w \in \mathbb{R}^d : \exists j \in [d_i], w_j^i \neq 0\}$ , and  $m > 0$  is the maximum allowable aggregate penalty. The choice of the partition can be made to simplify a neural network's architecture, for example each  $w^i$  can be the weights and bias of a neuron in a fully connected layer or of a filter in a convolutional layer. If  $m$  is an upper bound on the available memory to store  $w$  on a device, then each  $p_i$  can be the amount of memory required for each  $w^i$ . We assume that each  $p_i \leq m$ . If there exists a  $p_i > m$ , the associated decision variables  $w^i$  can be removed without affecting problem (1). For the Euclidean projection operator  $\Pi_C(\cdot)$  to be nonempty, it is sufficient that  $C$  is a closed set (Rockafellar and Wets, 2009, Example 1.20), which is verified in the next proposition.

**Proposition 1**  $C$  is a closed set.

### 4. Related works

The projection and normal cone of  $C$  are presented in Section 6, which is an extension of the analysis of the  $l_0$ -norm constraint in (Bauschke et al., 2014). Non-asymptotic convergence to an expected  $\epsilon$ -stationary point has been established for the proximal mini-batch SGD algorithm under the assumption that the function  $f$  has a Lipschitz continuous gradient in (Xu et al., 2019). In general, neural networks are not differentiable so this result cannot be applied. More appropriate for deep learning optimization is the assumption that  $f(w)$  is

(locally) Lipschitz continuous.

For asymptotic convergence results, in (Davis et al., 2020), the stochastic subgradient algorithm is proven to converge asymptotically to a Clarke stationary point almost surely for locally Lipschitz functions which admit a Whitney stratifiable graph, for step-sizes approaching zero in the limit, with an extension to the proximal stochastic subgradient algorithm. In (Bianchi et al., 2022), the authors consider a fixed step-size and model the randomness of stochastic gradients in a manner more congruent with using SGD for locally Lipschitz loss functions, and prove a convergence result in probability to the set of Clarke stationary points. The authors also consider a projected SGD algorithm, in particular for closed Euclidean balls, which ameliorates some technical assumptions. A locally Lipschitz continuous generalized-differentiable (Norkin, 1980) function with a convex and closed constraint is considered in (Ruszczynski, 2020). Asymptotic convergence to a Clarke stationary point for a stochastic subgradient method with averaging is proven.

Non-asymptotic convergence for a zeroth-order algorithm is presented in (Nesterov and Spokoiny, 2017) for the minimization of deterministic Lipschitz continuous functions using Gaussian smoothing. Non-asymptotic convergence results for first-order methods in terms of the Clarke  $\epsilon$ -subdifferential, in the deterministic and stochastic setting are given in (Zhang et al., 2020), under the assumption that loss functions are directionally-differentiable, and in the stochastic setting in (Metel and Takeda, 2022) using iterate perturbation. A comparison of our convergence criteria and computational complexity is given in Section 7.

## 5. Randomized smoothing of $f(w)$

To overcome the non-differentiability of  $f(w)$ , the original problem can be replaced by a smoothed approximation (see Proposition 6),

$$\min_{w \in \mathbb{R}^d} f_\alpha(w) \quad \text{s.t. } w \in C \cap \overline{B}_\beta,$$

where  $f_\alpha(w) := \mathbb{E}[f(w + u)]$  for a random vector  $u : \Omega \rightarrow \mathbb{R}^d$  uniformly distributed over  $\overline{B}_{\frac{\alpha}{2}}$  for an  $\alpha > 0$ . All  $u_i$  are mutually independent random variables with marginal probability distributions equal to

$$P_{u_i} = \begin{cases} \frac{1}{\alpha} & \text{if } |u_i| \leq \frac{\alpha}{2} \\ 0 & \text{otherwise.} \end{cases}$$

The added constraint  $\overline{B}_\beta$  for a  $\beta > 0$  is to allow us to assume that  $f(w)$  is only Lipschitz continuous over a compact set around zero. If  $f(w)$  is Lipschitz continuous over  $\mathbb{R}^d$ , this constraint can be removed by setting  $\beta = \infty$ . The assumptions on  $F(w, \xi)$  are similar to those used in (Metel and Takeda, 2022). For a  $\kappa > \beta + \frac{\alpha}{2}$ , we assume that  $F(w, \xi)$  is a  $\mathcal{B}_{\overline{B}_\kappa \times \mathbb{R}^p}$ -measurable function, where  $\mathcal{B}_{(\cdot)}$  denotes the Borel  $\sigma$ -algebra. We assume that for each  $\xi \in \mathbb{R}^p$ ,  $F(w, \xi)$  is continuous in  $w \in \overline{B}_\kappa$ , and for a measurable function  $L_0(\xi)$ ,  $F(w, \xi)$  is  $L_0(\xi)$ -Lipschitz continuous,

$$|F(w, \xi) - F(w', \xi)| \leq L_0(\xi) \|w - w'\|_2, \quad (4)$$

for all  $w, w' \in \overline{B}_\kappa$  and for all  $\xi \in \mathbb{R}^p$  outside of a Borel null set. It is assumed that  $L_0(\xi)$  is square integrable,  $Q := \mathbb{E}[L_0(\xi)^2] < \infty$ . It follows that  $f(w)$  is Lipschitz continuous in  $w \in \overline{B}_\kappa$ .

**Proposition 2** *The function  $f$  is  $L_0 := \mathbb{E}[L_0(\xi)]$ -Lipschitz continuous over  $\overline{B}_\kappa$ .*

Given that  $f(w)$  is Lipschitz continuous over  $w \in \overline{B}_\kappa$ , it is differentiable almost everywhere over  $w \in \overline{B}_{\beta+\alpha/2}$  by Rademacher's theorem (Heinonen, 2004, Theorem 3.1). The function  $\nabla f$  may not be defined on a null set in  $\overline{B}_{\beta+\alpha/2}$ , so we define  $\widetilde{\nabla} f(w)$  to be a  $\mathcal{B}_{\overline{B}_{\beta+\alpha/2}}$ -measurable function which for every  $w \in \overline{B}_\beta$  equals  $\nabla f(w + u)$  for almost every  $u$  over  $(\overline{B}_{\alpha/2}, \mathcal{B}_{\overline{B}_{\alpha/2}}, P_u)$ . Similarly, the function  $F(w, \xi)$  is differentiable almost everywhere over the product measure space  $(\overline{B}_{\beta+\alpha/2} \times \mathbb{R}^p, \mathcal{B}_{\overline{B}_{\beta+\alpha/2} \times \mathbb{R}^p}, m \times P_\xi)$ , where  $m$  is the Lebesgue measure restricted to Borel sets (Metel and Takeda, 2022, Property 1). We define  $\widetilde{\nabla} F(w, \xi)$  to be a  $\mathcal{B}_{\overline{B}_{\beta+\alpha/2} \times \mathbb{R}^p}$ -measurable function, which for every  $w \in \overline{B}_\beta$  equals  $\nabla F(w + u, \xi)$  for almost every  $(u, \xi)$  over  $(\overline{B}_{\alpha/2} \times \mathbb{R}^p, \mathcal{B}_{\overline{B}_{\alpha/2} \times \mathbb{R}^p}, P_u \times P_\xi)$ . Applying the results of (Bolte and Pauwels, 2021), it is verified in Section 8 that the output of backpropagation has the key properties of  $\widetilde{\nabla} F(w, \xi)$ , namely measurability and being equal to  $\nabla F(w, \xi)$  almost everywhere for conditions which are widely applicable for deep learning applications.

Another approach to overcome the non-differentiability of  $f(w)$  is to consider a zeroth-order algorithm. As proposed in (Gupal, 1977), an unbiased stochastic estimation of the gradient of  $f_\alpha(w)$  can be computed using the following finite-difference functions,  $df : \mathbb{R}^{2d} \rightarrow \mathbb{R}^d$  and  $dF : \mathbb{R}^{2d+p} \rightarrow \mathbb{R}^d$ , defined component-wise as

$$\begin{aligned} df_i(w, u_{\setminus i}) &:= f(w_1 + u_1, \dots, w_{i-1} + u_{i-1}, w_i + \frac{\alpha}{2}, w_{i+1} + u_{i+1}, \dots, w_d + u_d) \\ &\quad - f(w_1 + u_1, \dots, w_{i-1} + u_{i-1}, w_i - \frac{\alpha}{2}, w_{i+1} + u_{i+1}, \dots, w_d + u_d), \end{aligned}$$

and

$$\begin{aligned} dF_i(w, u_{\setminus i}, \xi) &:= F(w_1 + u_1, \dots, w_{i-1} + u_{i-1}, w_i + \frac{\alpha}{2}, w_{i+1} + u_{i+1}, \dots, w_d + u_d, \xi) \\ &\quad - F(w_1 + u_1, \dots, w_{i-1} + u_{i-1}, w_i - \frac{\alpha}{2}, w_{i+1} + u_{i+1}, \dots, w_d + u_d, \xi), \end{aligned}$$

where  $u_{\setminus i} := [u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_d]^T$ .<sup>1</sup>

### 5.1 Properties of $f_\alpha(w)$

Given the assumptions made about the use of randomized smoothing and the stochastic function  $F(w, \xi)$ , the following are resulting properties of  $f_\alpha(w)$  which will be useful in Section 7 for the analysis of the proposed training algorithms. For some similar results when  $f(w)$  is Lipschitz continuous over  $\mathbb{R}^d$  and the random vector  $u$  is Gaussian, see (Nesterov and Spokoiny, 2017). Gupal (1977) motivated the use of uniform perturbation over  $l_\infty$ -balls,

1. We use the notation  $df(w, u)$  and  $dF(w, u, \xi)$ , but then switch to  $df_i(w, u_{\setminus i})$  and  $dF_i(w, u_{\setminus i}, \xi)$  to make it clear that there is no  $u_i$  argument for the  $i^{\text{th}}$  component function.

where similar results to Propositions 3 and 6 can be found.

The following proposition proves that unbiased estimates of  $\nabla f_\alpha(w)$  can be generated using  $\tilde{\nabla}f$ ,  $\tilde{\nabla}F$ ,  $df$ , or  $dF$  with samples of  $u$  and  $\xi$ .

**Proposition 3** For all  $w \in \overline{B}_\beta$ ,

$$\begin{aligned}\nabla f_\alpha(w) &= \mathbb{E}[\tilde{\nabla}f(w+u)] = \mathbb{E}[\tilde{\nabla}F(w+u, \xi)] \\ &= \alpha^{-1}\mathbb{E}[df(w, u)] = \alpha^{-1}\mathbb{E}[dF(w, u, \xi)].\end{aligned}$$

The next proposition relates the gradient of  $f_\alpha(w)$  with the Clarke  $\epsilon$ -subdifferential of  $f(w)$ .

**Proposition 4** Assume that  $\kappa > \beta + \sqrt{d}\frac{\alpha}{2}$ . For all  $w \in \overline{B}_\beta$  with  $\hat{\alpha} = \sqrt{d}\frac{\alpha}{2}$ ,  $\nabla f_\alpha(w) \in \overline{\partial}_{\hat{\alpha}}f(w)$ .

Proposition 4 required a stronger condition on  $\kappa$  to ensure that  $\overline{\partial}_{\hat{\alpha}}f(w)$  is well-defined, meaning that the Clarke subdifferential is only being considered for values of  $x \in \mathbb{R}^d$  where  $f(w)$  is Lipschitz continuous on a neighbourhood of  $x$ . The following proposition focuses on properties of  $f_\alpha(w)$  and its relation to  $f(w)$ .

**Proposition 5**

1.  $f_\alpha(w)$  is  $L_0$ -Lipschitz continuous for  $w \in \overline{B}_\beta$ .
2. For all  $w \in \overline{B}_\beta$ ,  $|f_\alpha(w) - f(w)| \leq \alpha L_0 \sqrt{\frac{d}{12}}$ .
3. For a closed set  $S$ , let  $w_\alpha^*$  and  $w^*$  be minimizers of  $f_\alpha(w)$  and  $f(w)$  respectively for  $w \in S \cap \overline{B}_\beta$ , then  $|f_\alpha(w_\alpha^*) - f(w^*)| \leq \alpha L_0 \sqrt{\frac{d}{12}}$ .
4. For any two values  $w, w' \in \overline{B}_\beta$ ,  $|f_\alpha(w) - f_\alpha(w')| \leq 2\beta\sqrt{d}L_0$ .

The following proposition gives the Lipschitz constant of  $\nabla f_\alpha(w)$ , and will be referred to as the smoothness of  $f_\alpha(w)$ .

**Proposition 6** For all  $w \in \overline{B}_\beta$ ,  $\nabla f_\alpha(w)$  is  $2\alpha^{-1}\sqrt{d}L_0$ -Lipschitz continuous.

Considering the sample mean of a mini-batch of estimators of  $\nabla f_\alpha(w)$ , the next proposition gives bounds on the trace of their covariance matrices and on the expected value of their squared  $l_2$ -norm, which will be used in the convergence analysis of Section 7.

**Proposition 7** For all  $w \in \overline{B}_\beta$ ,

1.  $\mathbb{E}[\|\nabla f_\alpha(w) - \frac{1}{M\alpha} \sum_{i=1}^M dF(w, u^i, \xi^i)\|_2^2] \leq \frac{dQ}{M}$
2.  $\mathbb{E}[\|\nabla f_\alpha(w) - \frac{1}{M} \sum_{i=1}^M \tilde{\nabla}F(w+u^i, \xi^i)\|_2^2] \leq \frac{Q}{M}$
3.  $\mathbb{E}[\|\frac{1}{M\alpha} \sum_{i=1}^M dF(w, u^i, \xi^i)\|_2^2] \leq dQ$
4.  $\mathbb{E}[\|\frac{1}{M} \sum_{i=1}^M \tilde{\nabla}F(w+u^i, \xi^i)\|_2^2] \leq Q,$

where  $\{u^i\}$  and  $\{\xi^i\}$  are independent samples of  $u$  and  $\xi$ .

## 6. Properties of $C \cap \overline{B}_\beta$

In this section we give the projection onto  $C \cap \overline{B}_\beta$ , and its Fréchet and Mordukhovich normal cones. For the projection and normal cone of the set  $\{w \in \mathbb{R}^d : \|w\|_0 \leq m\}$ , see (Bauschke et al., 2014). The projection onto  $C \cap \overline{B}_\beta$  requires solving a 0-1 knapsack problem. This problem is NP-complete, though it can be solved in pseudo-polynomial time when all  $p_i \in \mathbb{Z}_{>0}$  and  $m \in \mathbb{Z}_{>0}$ , which holds when allocating memory as described in Section 3. For further background on this problem and algorithms see (Kellerer et al., 2004).

### 6.1 Projection onto $C \cap \overline{B}_\beta$

The next proposition shows how the projection onto  $C \cap \overline{B}_\beta$  can be computed using a 0-1 knapsack problem.

**Proposition 8** *For any  $w \in \mathbb{R}^d$ , let  $Z^*$  equal the set of optimal solutions of the 0-1 knapsack problem,*

$$\begin{aligned} \max_{z \in \{0,1\}^n} \quad & \sum_{i=1}^n z_i (\|w^i\|_2^2 - \|\max(|w^i| - \beta, 0)\|_2^2) \\ \text{s.t.} \quad & \sum_{i=1}^n z_i p_i \leq m, \end{aligned} \tag{5}$$

where  $\|\max(|w^i| - \beta, 0)\|_2^2 := \sum_{j=1}^{d_i} (\max(|w_j^i| - \beta, 0))^2$ . The projection  $\Pi_{C \cap \overline{B}_\beta}(w)$  equals

$$\begin{aligned} \Pi_{C \cap \overline{B}_\beta}(w) = \{x \in \mathbb{R}^d : \exists z^* \in Z^*, \quad & x^i = \text{sgn}(w^i) \min(|w^i|, \beta) \text{ if } z_i^* = 1, \\ & x^i = 0 \text{ otherwise } \forall i \in [n]\}, \end{aligned}$$

where  $x^i = \text{sgn}(w^i) \min(|w^i|, \beta)$  denotes  $x_j^i = \text{sgn}(w_j^i) \min(|w_j^i|, \beta)$  for  $j = 1, 2, \dots, d_i$ .

The following remark gives the projection onto  $C$ , which can be verified by taking  $\beta \rightarrow \infty$  in Proposition 8.

**Remark 9** *For the projection onto  $C$  for any  $w \in \mathbb{R}^d$ , the 0-1 knapsack problem (5) becomes*

$$\begin{aligned} \max_{z \in \{0,1\}^n} \quad & \sum_{i=1}^n z_i \|w^i\|_2^2 \\ \text{s.t.} \quad & \sum_{i=1}^n z_i p_i \leq m. \end{aligned} \tag{6}$$

If  $Z^*$  equals the set of optimal solutions of (6), then the projection  $\Pi_C(w)$  equals

$$\Pi_C(w) = \{x \in \mathbb{R}^d : \exists z^* \in Z^*, \quad x^i = w^i \text{ if } z_i^* = 1, \quad x^i = 0 \text{ otherwise } \forall i \in [n]\}.$$



## 6.2 Normal cones of $C \cap \overline{B}_\beta$

Assume that  $w \in C \cap \overline{B}_\beta$ , and let  $I(w) := \{i \in [n] : w^i \neq 0\}$  be the indices of the subsets of non-zero weights, and let  $J(w) := \{j \in [n] \setminus I(w) : \sum_{i \in I(w)} p_i + p_j \leq m\}$  be the indices of subsets which are zero, but are not constrained to be. The following proposition gives the Fréchet normal cone to the set  $C \cap \overline{B}_\beta$ .

**Proposition 10** *For any  $w \in C \cap \overline{B}_\beta$ ,*

$$\widehat{N}(w, C \cap \overline{B}_\beta) = \left\{ y \in \mathbb{R}^d : \forall i \in I(w) \cup J(w), \forall j \in [d_i], y_j^i \in \begin{cases} \mathbb{R}_{\geq 0} & \text{if } w_j^i = \beta \\ 0 & \text{if } |w_j^i| < \beta \\ \mathbb{R}_{\leq 0} & \text{if } w_j^i = -\beta \end{cases} \right\}. \quad (7)$$

The following remark gives the Fréchet normal cone to the set  $C$ , which can be verified by taking  $\beta \rightarrow \infty$  in Proposition 10.

**Remark 11** *For any  $w \in C$ ,*

$$\widehat{N}(w, C) = \{y \in \mathbb{R}^d : \forall i \in I(w) \cup J(w), y^i = 0\}.$$

Let  $Y := \{X \subseteq 2^{[n]} : I(w) \subseteq X, \sum_{i \in X} p_i \leq m \text{ and } \sum_{i \in X} p_i + p_j > m \forall j \notin X\}$ , which contains all of the sets of indices  $X$  containing  $I(w)$  which make the constraint  $\sum_{i \in X} p_i \leq m$  tight, in the sense that no further feasible index can be added to  $X$ . The next proposition gives the Mordukhovich normal cone to the set  $C \cap \overline{B}_\beta$ .

**Proposition 12** *For any  $w \in C \cap \overline{B}_\beta$ ,*

$$N(w, C \cap \overline{B}_\beta) = \left\{ y \in \mathbb{R}^d : \exists X \in Y, \forall i \in X, \forall j \in [d_i], y_j^i \in \begin{cases} \mathbb{R}_{\geq 0} & \text{if } w_j^i = \beta \\ 0 & \text{if } |w_j^i| < \beta \\ \mathbb{R}_{\leq 0} & \text{if } w_j^i = -\beta \end{cases} \right\}. \quad (8)$$

The next remark gives the Mordukhovich normal cone to the set  $C$ .

**Remark 13** *For any  $w \in C$ ,*

$$N(w, C) = \{y \in \mathbb{R}^d : \exists X \in Y, y^i = 0 \forall i \in X\}.$$

## 7. Training Algorithm

---

**Algorithm 1** Stochastic Projected Algorithm (SPA)

---

**Input:**  $w^1 \in S \cap \overline{B}_\beta$ ,  $\eta > 0$ ,  $K \in \mathbb{Z}_{>0}$ ,  $M \in \mathbb{Z}_{>0}$   
 $R \sim \text{uniform}\{2, \dots, K + 1\}$   
**for**  $k = 1, 2, \dots, R - 1$  **do**  
    Sample  $u^{k,i} \sim P_u$  for  $i = 1, \dots, M$   
    Sample  $\xi^{k,i} \sim P_\xi$  for  $i = 1, \dots, M$   
    (1)  $w^{k+1} \in \Pi_{S \cap \overline{B}_\beta}(w^k - \frac{\eta}{M\alpha} \sum_{i=1}^M dF(w^k, u^{k,i}, \xi^{k,i}))$   
    OR  
    (2)  $w^{k+1} \in \Pi_{S \cap \overline{B}_\beta}(w^k - \frac{\eta}{M} \sum_{i=1}^M \tilde{\nabla} F(w^k + u^{k,i}, \xi^{k,i}))$   
**end for**  
**Output:**  $w^R$

---

The following convergence results of SPA (Algorithm 1) are applicable for any constraint set  $S$  which is closed and for which there exists a computable element of the Euclidean projection onto  $S \cap \overline{B}_\beta$ . If (4) holds for all  $w, w' \in \mathbb{R}^d$ , then the projection operator can be simplified to  $\Pi_S$ . The proof of Theorem 15 is an adaptation of the proof of (Xu et al., 2019, Theorem 2). For some similar results of this section for a first-order algorithm for unconstrained problems see (Metel and Takeda, 2022). SPA has two settings: (1) is a zeroth-order and (2) is a first-order algorithm. The algorithm requires that  $w^1 \in S \cap \overline{B}_\beta$ . A simple choice is to pick an arbitrary  $w^0 \in \mathbb{R}^d$  and to set  $w^1 \in \Pi_{S \cap \overline{B}_\beta}(w_0)$ , but when training a neural network with  $S = C$  and a high sparsity level, there is a risk of initializing the neural network with layer collapse (Hoeffler et al., 2021, Page 20), where all weights in a layer are set to zero, disconnecting the network. We highlight that the initial  $w^1 \in S \cap \overline{B}_\beta$  can be chosen to ensure that there are non-zero weights in each layer, or any other desired property.

We consider two convergence criteria. A solution  $\bar{w}$  is an expected  $(\epsilon_1, \epsilon_2)$ -stationary point if

$$|f_\alpha(w) - f(w)| \leq \epsilon_1 \text{ for all } w \in \overline{B}_\beta, \quad \text{and} \quad \mathbb{E}[\text{dist}(0, \nabla f_\alpha(\bar{w}) + N(\bar{w}, S \cap \overline{B}_\beta))] \leq \epsilon_2,$$

which guarantees that  $\bar{w}$  is an expected  $\epsilon_2$ -stationary point for a smooth approximation of  $f(w)$  with a uniform error from  $f(w)$  within  $\epsilon_1$ . A solution  $\bar{w}$  is an expected  $(\epsilon_3, \epsilon_4)$ -stationary point if

$$\hat{\alpha} \leq \epsilon_3 \quad \text{and} \quad \mathbb{E}[\text{dist}(0, \bar{\partial} f_{\hat{\alpha}}(\bar{w}) + N(\bar{w}, S \cap \overline{B}_\beta))] \leq \epsilon_4,$$

which can be seen as a relaxation of an expected  $\epsilon_4$ -stationary point, replacing the Mordukhovich with a Clarke  $\hat{\alpha}$ -subdifferential. It will also be used in Theorem 20 for a method with an asymptotic convergence guarantee to a C-M stationary point. These convergence criteria are related as for sufficiently small  $\alpha$  an  $(\epsilon_1, \epsilon_2)$ -stationary point implies an  $(\epsilon_3, \epsilon_4)$ -stationary point with  $\epsilon_3 = \sqrt{d} \frac{\alpha}{2}$  and  $\epsilon_4 = \epsilon_2$  using Proposition 4.

The next proposition verifies the existence of a Borel measurable selection of the projection operator  $\Pi_{S \cap \bar{B}_\beta}(\cdot)$ , and of the measurability of the distance functions used in the convergence criteria. By the assumptions that  $F(w, \xi)$  and  $\tilde{\nabla}F(w, \xi)$  are Borel measurable, the iterates  $\{w^k\}$  from SPA are measurable using such a selection of  $\Pi_{S \cap \bar{B}_\beta}(\cdot)$ . This proposition also covers C-M stationary points, i.e.  $\bar{\partial}_0 f(w) = \bar{\partial} f(w)$  (Goldstein, 1977, Corollary 2.5).

**Proposition 14** *Let  $S$  be a closed set. There exists a measurable selection of  $\Pi_{S \cap \bar{B}_\beta}(\cdot)$ . For any  $0 \leq \epsilon \leq \frac{\alpha}{2} \text{dist}(0, \bar{\partial}_\epsilon f(w) + N(w, S \cap \bar{B}_\beta))$ , and  $\text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \bar{B}_\beta))$  are Borel measurable functions in  $w \in \bar{B}_\beta$ .*

Together with Proposition 5.2, the following theorem presents the non-asymptotic convergence to an expected  $(\epsilon_1, \epsilon_2)$ -stationary point of SPA.

**Theorem 15** *Let  $\eta = \frac{\alpha}{3\rho\sqrt{d}L_0}$  for  $\rho > 0$  and let  $\tau \geq 0$  such that  $\rho + \tau > 1$ . For a solution from SPA given any choice of  $w^1 \in S \cap \bar{B}_\beta$ ,  $K \in \mathbb{Z}_{>0}$ , and  $M \in \mathbb{Z}_{>0}$ , it holds that*

$$\mathbb{E}[\text{dist}(0, \nabla f_\alpha(w^R) + N(w^R, S \cap \bar{B}_\beta))^2] \leq C_1 \frac{2\alpha^{-1}\sqrt{d}L_0\Delta}{K} + C_2 \frac{vQ}{M}, \quad (9)$$

where

$$\begin{aligned} \Delta &\geq f_\alpha(w^1) - f_\alpha(w_\alpha^*) \\ &\leq \min \left( 2\beta\sqrt{d}L_0, f(w^1) - f(w^*) + \alpha L_0 \sqrt{\frac{d}{3}} \right) \end{aligned} \quad (10)$$

for minimizers  $w_\alpha^*$  and  $w^*$  of  $f_\alpha(w)$  and  $f(w)$  respectively for  $w \in S \cap \bar{B}_\beta$ ,  $C_1 := \frac{2(1+3\rho)}{(\tau+\rho-1)} + 3\rho$ ,

$$C_2 := \frac{4(1+3\rho)}{(\tau+\rho-1)} \left( \frac{1}{2} + \frac{2}{3} \frac{M\tau}{\rho^2} \right) + 3, \text{ and } v := \begin{cases} d & \text{if using (1)} \\ 1 & \text{if using (2)}. \end{cases}$$

Inequality (10) gives valid choices for  $\Delta$  which are easier to compute and related to the true loss function  $f$ . Using Theorem 15 and Proposition 4, the following corollary holds.

**Corollary 16** *Assume that  $\kappa > \beta + \sqrt{d}\frac{\alpha}{2}$  and  $\hat{\alpha} = \sqrt{d}\frac{\alpha}{2}$ . Let  $\eta$ ,  $\rho$ ,  $\tau$ ,  $\Delta$ ,  $C_1$ , and  $C_2$  be defined as in Theorem 15. For a solution from SPA given any choice of  $w^1 \in S \cap \bar{B}_\beta$ ,  $K \in \mathbb{Z}_{>0}$ , and  $M \in \mathbb{Z}_{>0}$ , it holds that*

$$\mathbb{E}[\text{dist}(0, \bar{\partial} f_{\hat{\alpha}}(w^R) + N(w^R, S \cap \bar{B}_\beta))^2] \leq C_1 \frac{2\alpha^{-1}\sqrt{d}L_0\Delta}{K} + C_2 \frac{vQ}{M}. \quad (11)$$

The parameter  $\tau$  is not required and in fact  $\tau > 0$  results in the constant term

$$\frac{4(1+3\rho)}{(\tau+\rho-1)} \left( \frac{2}{3} \frac{\tau}{\rho^2} \right) vQ$$

in the expansion of the right-hand-side of (9) and (11). The parameter  $\tau$  is included so that the convergence bounds are applicable for any step-size  $\eta > 0$ , though it will likely be

poor unless  $Q \approx 0$ . A large  $\sqrt{d}$ , such as for deep neural networks, will result in a small step-size  $\eta$  as well as require a large  $K$  to get an adequate convergence guarantee. The inclusion of  $\tau$  is also an attempt to remedy this when  $Q \approx 0$ , i.e. replacing  $\rho > 0$  and  $\tau = 0$  with  $\rho' > 0$  and  $\tau' > 0$  such that  $\rho' + \tau' = \rho$  will increase  $\eta$  and decrease  $C_1$ . For the remainder of this section we will assume that  $\tau = 0$ , which results in  $C_1$  and  $C_2$  being equal to  $C_1^{\tau=0} := \frac{2+3\rho+3\rho^2}{\rho-1}$  and  $C_2^{\tau=0} := \frac{9\rho-1}{\rho-1}$ . Table 1 presents some choices for  $\rho$  resulting in  $C_1^{\tau=0}$  and  $C_2^{\tau=0}$  being integer-valued.

Table 1: Some choices for  $\rho$ .

| $\rho$         | $\frac{4}{3}$ | $\frac{5}{3}$ | 2  | 5  |
|----------------|---------------|---------------|----|----|
| $C_1^{\tau=0}$ | 34            | 23            | 20 | 23 |
| $C_2^{\tau=0}$ | 33            | 21            | 17 | 11 |

The following corollary gives the computational complexity to guarantee an expected  $(\epsilon_1, \epsilon_2)$  or  $(\epsilon_3, \epsilon_4)$ -stationary point in terms of the number of either  $dF(w^k, u^{k,i}, \xi^{k,i})$  or  $\nabla F(w^k + u^{k,i}, \xi^{k,i})$  computations, which will be referred to as *gradient calls*, and in terms of the number of projections.

**Corollary 17** *Running SPA as described in Theorem 15 with  $\alpha = \frac{\epsilon_1}{L_0\sqrt{\frac{d}{12}}}$ ,  $\tau = 0$ ,*

$$K = \left\lceil C_1 \sqrt{\frac{4}{3}} \frac{dL_0^2\Delta}{\epsilon_1\epsilon_2^2} \right\rceil, \quad \text{and} \quad M = \left\lceil C_2 \frac{2vQ}{\epsilon_2^2} \right\rceil$$

*guarantees an expected  $(\epsilon_1, \epsilon_2)$ -stationary point. Assuming that  $\kappa > \beta + \epsilon_3$ , and setting  $\alpha = \frac{2\epsilon_3}{\sqrt{d}}$ ,  $\tau = 0$ ,*

$$K = \left\lceil C_1 \frac{2dL_0\Delta}{\epsilon_3\epsilon_4^2} \right\rceil, \quad \text{and} \quad M = \left\lceil C_2 \frac{2vQ}{\epsilon_4^2} \right\rceil$$

*guarantees an expected  $(\epsilon_3, \epsilon_4)$ -stationary point. Using these choices of  $\alpha$ ,  $\tau$ ,  $K$ , and  $M$  give gradient call complexities of  $O(\epsilon_1^{-1}\epsilon_2^{-4})$  and  $O(\epsilon_3^{-1}\epsilon_4^{-4})$ , and projection operator complexities of  $O(\epsilon_1^{-1}\epsilon_2^{-2})$  and  $O(\epsilon_3^{-1}\epsilon_4^{-2})$  to achieve an expected  $(\epsilon_1, \epsilon_2)$  and  $(\epsilon_3, \epsilon_4)$ -stationary point, respectively.*

The next corollary gives the computational complexity for an  $(\epsilon_1, \epsilon_2)$  or  $(\epsilon_3, \epsilon_4)$ -stationary point with a probability of at least  $1 - \gamma$  for any  $\gamma \in (0, 1)$  using the method proposed in (Ghadimi and Lan, 2013, Section 2.2). SPA is required to be run  $r \in \mathbb{Z}_{>0}$  times, generating  $r$  different solutions, with the result holding for the solution which minimizes the distance to stationarity using a sample mean approximation of  $\nabla f_\alpha(w)$ .

**Corollary 18** *For any  $\gamma \in (0, 1)$  and  $\epsilon_1, \epsilon_2 > 0$  or  $\epsilon_3, \epsilon_4 > 0$ , with  $\kappa > \beta + \epsilon_3$ , assume that SPA is run  $r := \lceil -\ln(c\gamma) \rceil$  times for any  $c \in (0, 1)$  according to Theorem 15 with  $\tau = 0$ ,*

$$\alpha = \frac{\epsilon_1}{L_0 \sqrt{\frac{d}{12}}} \text{ or } \alpha = \frac{2\epsilon_3}{\sqrt{d}},$$

$$K = \left[ C_1 \sqrt{\frac{4}{3}} \frac{dL_0^2 \Delta}{\epsilon_1 (\epsilon'_2)^2} \right] \quad \text{and} \quad M = \left[ C_2 \frac{2\nu Q}{(\epsilon'_2)^2} \right], \quad \text{or}$$

$$K = \left[ C_1 \frac{2dL_0 \Delta}{\epsilon_3 (\epsilon'_4)^2} \right] \quad \text{and} \quad M = \left[ C_2 \frac{2\nu Q}{(\epsilon'_4)^2} \right],$$

where  $\epsilon'_2 = \sqrt{\frac{\epsilon_2^2 - 6\psi \frac{Q}{T}}{4e}}$ ,  $\epsilon'_4 = \sqrt{\frac{\epsilon_4^2 - 6\psi \frac{Q}{T}}{4e}}$ ,  $\psi = \frac{[-\ln(c\gamma)]}{(1-c)\gamma}$ ,  $e := \exp(1)$ , and  $T = \lceil 6\phi\psi \frac{Q}{\epsilon_2^2} \rceil$  or  $T = \lceil 6\phi\psi \frac{Q}{\epsilon_4^2} \rceil$  for any  $\phi > 1$ , outputting solutions  $W := \{w^1, \dots, w^r\}$ . Let  $\{u^i\}_{i=1}^T$  and  $\{\xi^i\}_{i=1}^T$  be independent samples of  $u$  and  $\xi$ , and let  $w^* \in W$  be chosen such that

$$w^* \in \underset{w \in W}{\operatorname{argmin}} \operatorname{dist}(0, G(w) + N(w, S \cap \bar{B}_\beta)), \quad (12)$$

where

$$G(w) := \begin{cases} \frac{1}{T\alpha} \sum_{i=1}^T dF(w, u^i, \xi^i) & \text{if using (1)} \\ \frac{1}{T} \sum_{i=1}^T \tilde{\nabla} F(w + u^i, \xi^i) & \text{if using (2)} \end{cases}$$

in SPA. It follows that  $w^*$  is an  $(\epsilon_1, \epsilon_2)$  or  $(\epsilon_3, \epsilon_4)$ -stationary point with a probability of at least  $1 - \gamma$ , it is generated with  $\tilde{O}(\epsilon_1^{-1}\epsilon_2^{-4} + \gamma^{-1}\epsilon_2^{-2})$  or  $\tilde{O}(\epsilon_3^{-1}\epsilon_4^{-4} + \gamma^{-1}\epsilon_4^{-2})$  gradient calls, and  $\tilde{O}(\epsilon_1^{-1}\epsilon_2^{-2})$  or  $\tilde{O}(\epsilon_3^{-1}\epsilon_4^{-2})$  projections.

The optimization problem (12) requires knowledge of the normal cone of  $S \cap \bar{B}_\beta$  as given in Proposition 12 for  $S = C$ , and is solved by computing the distance  $\operatorname{dist}(0, G(w) + N(w, S \cap \bar{B}_\beta))$   $\tilde{O}(1)$  times, once for each  $w \in W$ . The binary integer program discussed in the next proposition can be found in the proof, see (38), but its requirement is only for pathological cases in neural network training when a solution equals  $w^i = 0$  for an  $i \in [n]$  which is not constrained to be zero.

**Proposition 19** *If  $Y = I(w)$  as defined above Proposition 12,  $\operatorname{dist}(0, G(w) + N(w, C \cap \bar{B}_\beta)) = \|G(w) + v\|_2$  where*

$$v_j^i = \begin{cases} 0 & \text{if } i \in I(w) \text{ and } -U_j^i \\ -G_j^i(w) & \text{otherwise,} \end{cases}$$

where  $U_j^i := (|w_j^i| = \beta) \wedge (\operatorname{sgn}(G_j^i(w)) = -\operatorname{sgn}(w_j^i))$  for  $i \in [n]$  and  $j \in d_i$ . When there exists an  $X \in Y$  such that  $X \setminus I(w) \neq \{\emptyset\}$ , assume that  $\{p_i\} \subset \mathbb{Q}_{>0}$ . The distance  $\operatorname{dist}(0, G(w) + N(w, C \cap \bar{B}_\beta))$  can be computed by solving a binary integer program with  $|[n] \setminus I(w)|$  binary variables.

### Comparison of computational complexity

Our gradient call complexity matches that of (Nesterov and Spokoiny, 2017, Section 7) to achieve an expected  $(\epsilon_1, \epsilon_2)$ -stationary point for an unconstrained deterministic function  $f$ . Their random gradient-free oracle only requires two function evaluations, whereas  $df(w, u)$

requires  $2d$  function evaluations. The Gaussian smoothing is computationally appealing but we would need to assume that  $f(w)$  is Lipschitz continuous over  $\mathbb{R}^d$  as the function calls within  $df(w, u)$  would now be evaluated at any point in  $\mathbb{R}^d$  given the expanded image of normal random variables compared to  $u \in \overline{B}_{\frac{\sigma}{2}}$ . In (Metel and Takeda, 2022) the same computational complexity is proven for an expected  $(\epsilon_3, \epsilon_4)$ -stationary point for unconstrained stochastic functions. In (Zhang et al., 2020), a better computational complexity of  $\tilde{O}(\epsilon_3^{-1}\epsilon_4^{-3})$  is proven to achieve an  $(\epsilon_3, \epsilon_4)$ -stationary point in high probability in the deterministic setting. In the stochastic setting, (Zhang et al., 2020) proves a computational complexity of  $\tilde{O}(\epsilon_3^{-1}\epsilon_4^{-4})$  to achieve an expected  $(\epsilon_3, \epsilon_4)$ -stationary point similar to our work.

The next theorem proves that the set of solutions from running SPA with increasing accuracy has an asymptotic convergence guarantee to a C-M stationary point almost surely.

**Theorem 20** *Let  $\{\epsilon_3^i\}$  and  $\{\epsilon_4^i\}$  be strictly decreasing positive sequences approaching 0 in the limit, with  $\alpha^i$ ,  $K^i$  and  $M^i$  set to guarantee an expected  $(\epsilon_3^i, \epsilon_4^i)$ -stationary point running SPA according to Corollary 17 assuming that  $\kappa > \beta + \epsilon_3^1$ . Assume that SPA is run according to Theorem 15 with  $\tau = 0$ ,  $\alpha = \alpha^i$ ,  $K = K^i$ , and  $M = M^i$  for  $i = 1, 2, \dots$ , giving solutions  $\{w^i\}$ . If  $\beta$  is finite, there exists an accumulation point  $\bar{w}$  of  $\{w^i\}$  and it is a C-M stationary point almost surely. Otherwise, any accumulation point  $\bar{w}$  of  $\{w^i\}$  is a C-M stationary point almost surely.*

## 8. Using Backpropagation

This section considers computing  $\tilde{\nabla}F(w, \xi)$  using backpropagation for a problem setting entailing a wide range of deep learning applications. This is demonstrated using the results of (Bolte and Pauwels, 2021). Assume that  $\xi$  maps to a countable number of values  $\{\xi_i\}_{i=1}^\infty$  almost surely,  $\mathbb{P}(\xi \in \{\xi_i\}_{i=1}^\infty) = 1$ , and assume that for each  $\xi_k \in \{\xi_i\}_{i=1}^\infty$ ,  $F(w, \xi_k)$  for  $w \in \mathbb{R}^d$  can be written as a composition of locally Lipschitz continuous functions  $\{\sigma_j\}_{j \in I_k}$ , for an index set  $I_k$ , and assume that the functions  $\{\sigma_j\}_{j \in I_k}$  are definable in the same o-minimal structure.

**Proposition 21** (Bolte and Pauwels, 2021, Corollary 5) *For each  $\xi_k \in \{\xi_i\}_{i=1}^\infty$  set  $\tilde{\nabla}F(w, \xi_k)$  equal to the output of backpropagation using a measurable selection  $\tilde{\nabla}\sigma_j(\cdot) \in \bar{\partial}\sigma_j(\cdot)$ , which exists, for all  $j \in I_k$ , and for  $\xi \notin \{\xi_i\}_{i=1}^\infty$  set  $\tilde{\nabla}F(w, \xi) = a$  for any  $a \in \mathbb{R}^d$ .  $\tilde{\nabla}F(w, \xi)$  is Borel measurable for  $(w, \xi) \in \mathbb{R}^{d+p}$  and equals the gradient of  $F(w, \xi)$  for almost every  $(w, \xi) \in \mathbb{R}^{d+p}$ .*

We focus on the o-minimal structure of the ordered real exponential field,  $\mathbb{R}_{\text{exp}, <} := \mathbb{R}(+, \cdot, 0, 1, <, \text{exp})$ , which provides a wide class of definable functions typically found in deep learning architectures. For a short background on o-minimal structures see for example (Wilkie, 2007). The next proposition verifies the validity of using backpropagation for the building blocks used in the neural network considered in the next section, and also contains a sufficient background on o-minimal structures to understand the result. Other activation functions typically used in deep learning can be shown to have the following properties as well. We refer to what is computed during backpropagation as a *bp* gradient. Conv2d and MaxPool2d are defined as tensor-valued functions, but it is sufficient to consider their component functions separately.

Table 2: Parameters for MN and FMN datasets for  $s \in \{0.65, 0.5\}$ . All parameters were estimated or given explicitly by the theory of the paper given the choices for  $\kappa$ ,  $(\epsilon_1, \epsilon_2)$ , and  $\rho$ , except  $\beta = 0.99(\kappa - \alpha/2)$ , which satisfies  $\kappa > \beta + \frac{\alpha}{2}$ .

|     | $L_0$   | $Q$     | $\Delta$ | $\alpha$ | $\beta$ | $\eta$  | $K$    | $M$ |
|-----|---------|---------|----------|----------|---------|---------|--------|-----|
| MN  | 8.53E-2 | 7.49E-3 | 2.31     | 6.42E-2  | 1.66E-1 | 4.76E-4 | 4.38E5 | 2   |
| FMN | 1.09E-1 | 1.22E-2 | 2.30     | 5.05E-2  | 1.93E-1 | 2.68E-4 | 7.07E5 | 3   |

**Proposition 22** *The affine map, ReLU, the component functions of Conv2d and Max-Pool2d, and the loss function CrossEntropyLoss are definable in the o-minimal structure of  $\mathbb{R}_{exp,<}$ , and their bp gradients are measurable selections of their Clarke subdifferentials.*

## 9. Training a Neural Network

We trained a Lenet-5 type neural network on the MNIST (MN) and FashionMNIST (FMN) datasets constrained by  $C \cap \overline{B}_\beta$ . The projection operator  $\Pi_{C \cap \overline{B}_\beta}(\cdot)$  was computed using a branch-and-bound (BNB) algorithm to solve the 0 – 1 knapsack problem (5). A sampling approach was used to empirically estimate  $L_0$ ,  $Q$ , and  $\Delta$ . Details of the neural network architecture, the BNB algorithm, and the sampling approach can be found in Appendix F.

The constants  $L_0$  and  $Q$  are non-decreasing in  $\kappa$ , but it was observed that our estimates can be decreased significantly by decreasing  $\kappa$  without having much of an impact on training performance. This enabled reasonable choices for the required number of epochs implied by Corollary 17 with a choice of  $\kappa = 0.2$  and  $0.22$  for the MN and FMN datasets. We focused on the first-order version of SPA and ran it according to Corollary 17 to achieve an expected  $(\epsilon_1, \epsilon_2)$ -stationary point for  $\epsilon_1 = \epsilon_2 = 1/3$ . For the MN and FMN datasets,  $\rho$  was chosen as  $\rho = 2.5$  and  $2.75$  to minimize the required number of epochs, searching over a grid of 0.25 increments. We want to highlight that these parameters were chosen solely to ensure an adequate number of epochs, and similar or better solutions are expected for larger  $\kappa$ , smaller  $(\epsilon_1, \epsilon_2)$ , with reasonable values of  $\rho$ , e.g. within the domain of Table 1, but will require longer training times.

Weights and biases were grouped together by filter and neuron for the convolutional and fully connected layers to generate the partition  $\{w^i\}_{i=1}^n$ , where  $n = 236$ . The penalty  $p_i$  for  $i \in [n]$  was set to the dimension of each subset  $w^i$ ,  $p_i = d_i$ . The parameter  $m$  of  $C$  was chosen as  $m = (1 - s)d$  where  $s \in (0, 1)$  is the chosen sparsity level and  $d = 44426$ . Trying different values of  $s$  at 0.05 increments, layer collapse occurred with  $w_1 = \Pi_{C \cap \overline{B}_\beta}(w_0)$  for randomly initialized  $w_0$  and  $s = 0.7$ , so we restricted these experiments to  $s \leq 0.65$ . The values of  $\alpha$ ,  $K$ , and  $M$  were set according to Corollary 17, and  $\beta$  was set to  $\beta = 0.99(\kappa - \alpha/2)$ , such that  $\kappa > \beta + \frac{\alpha}{2}$  following Section 5. The value of  $\eta$  was set according to Theorem 15. Table 2 presents the values of the aforementioned estimated or computed parameters.

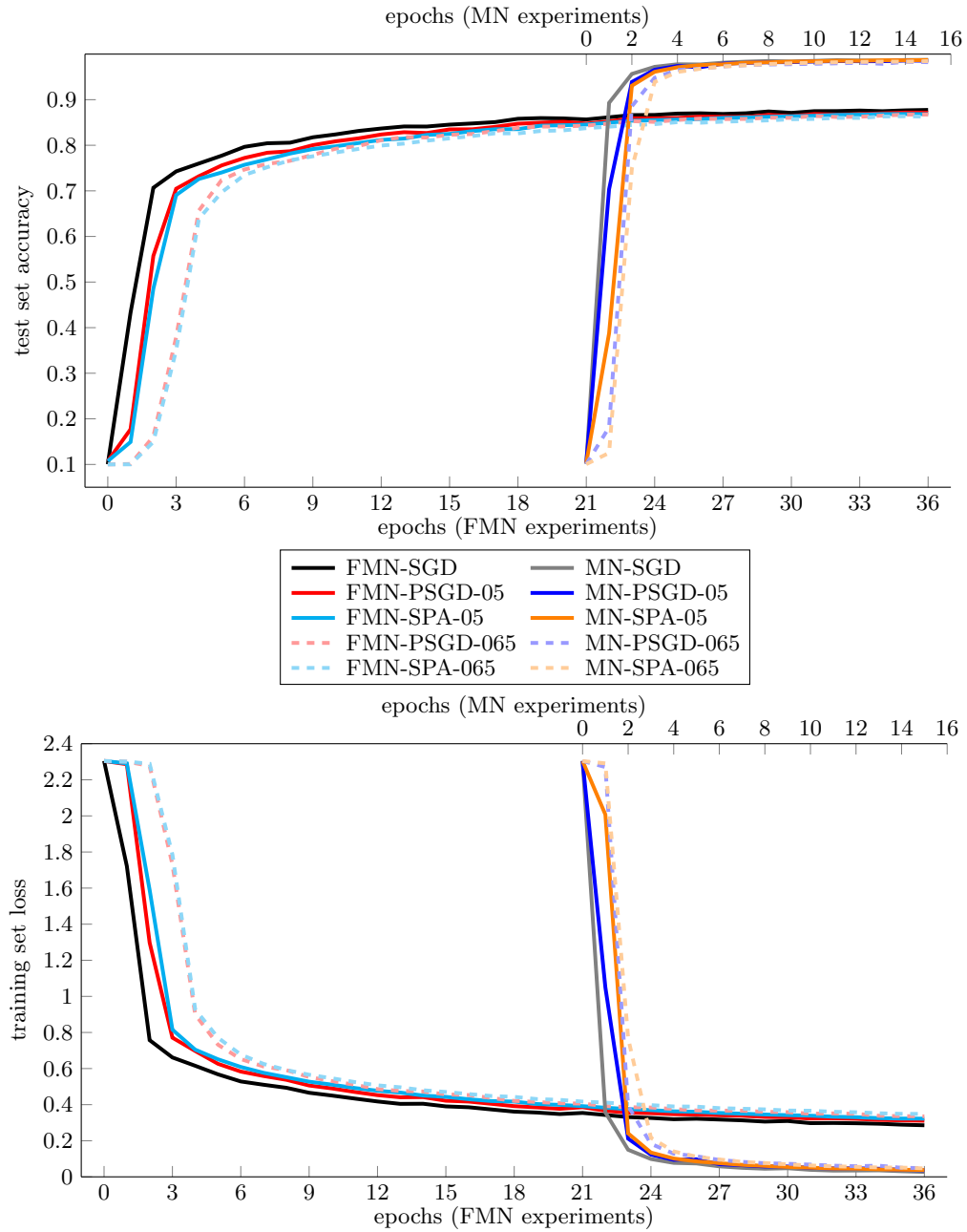


Figure 1: Test set accuracy and training set loss of PSGD and SPA. The numbers equal the sparsity level  $s$ .



SPA was compared to projected mini-batch SGD (PSGD) using the same parameterization but with no randomized smoothing, i.e.  $u^{k,i} = 0 \forall k, i$ , and unconstrained mini-batch SGD (SGD), run identically to PSGD but with no projection. All algorithms were run 3 times for  $\lceil KM/\mathcal{M} \rceil$  epochs, where  $\mathcal{M} = 60,000$  is the training set size, with the output averaged together. The experiments were run in Python 3.6.13 with Pytorch 1.8.1 on a server running Ubuntu 18.04.5 LTS with an Intel Xeon E5-2698 v4 CPU and an Nvidia Titan V GPU. Figure 1 plots the test set accuracy and the training set loss. The performance of SPA and PSGD are very similar, with better performance in earlier epochs with  $s = 0.5$ , and with all algorithms converging closely to SGD in later epochs.

The 0-1 knapsack problem did not pose a significant computational bottleneck to the training. An experiment measuring the computation time of  $\Pi_{C \cap \bar{B}_\beta}(\cdot)$  was conducted for the first 100 projections of 100 trials of the experimental setup of FMN-SGD-065. The projection was found to be the most challenging for  $w_0$ , with an average computation time of 0.168 seconds, with the remaining projections having an average computation time of 0.0790 seconds.

## 10. Conclusion

This paper studied theoretical aspects of structured sparsity for deep neural network training. A weighted group  $l_0$ -norm constraint was proposed and the projection operator and normal cone of this set were presented. The computational complexities of a zeroth and first-order stochastic projection algorithm for constrained Lipschitz continuous loss functions were given for  $(\epsilon_1, \epsilon_2)$  and  $(\epsilon_3, \epsilon_4)$ -stationary points in expectation and high probability, as well as a method with an asymptotic convergence guarantee to a C-M stationary point almost surely.

## Appendix A. Section 3 Proof

**Proof of Proposition 1** For a point  $x \in \{w^i \neq 0\}$ , choosing a  $j$  such that  $x_j^i \neq 0$ , it follows that  $B(x, |x_j^i|/2) \in \{w^i \neq 0\}$  as well, proving that  $\{w^i \neq 0\}$  is an open set. The lower level sets of  $p_i \mathbb{1}_{\{w^i \neq 0\}}(w)$ ,

$$\{w \in \mathbb{R}^d : p_i \mathbb{1}_{\{w^i \neq 0\}}(w) \leq \lambda\} = \begin{cases} \mathbb{R}^d & \text{for } p_i \leq \lambda \\ \{w^i \neq 0\}^c & \text{for } 0 \leq \lambda < p_i \\ \{\emptyset\} & \text{for } \lambda < 0 \end{cases}$$

are closed for all  $\lambda \in \mathbb{R}$ , which holds if and only if  $p_i \mathbb{1}_{\{w^i \neq 0\}}(w)$  is a lower semicontinuous function (Rockafellar and Wets, 2009, Theorem 1.6). A function  $h$  is lower semicontinuous at  $\bar{w} \in \mathbb{R}^d$  if  $\liminf_{w \rightarrow \bar{w}} h(w) \geq h(\bar{w})$  (Rockafellar and Wets, 2009, Definition 1.5). For all

$\bar{w} \in \mathbb{R}^d$ ,

$$\begin{aligned} \liminf_{w \rightarrow \bar{w}} \sum_{i=1}^n p_i \mathbb{1}_{\{w^i \neq 0\}}(w) &\geq \sum_{i=1}^n \liminf_{w \rightarrow \bar{w}} p_i \mathbb{1}_{\{w^i \neq 0\}}(w) \\ &\geq \sum_{i=1}^n p_i \mathbb{1}_{\{\bar{w}^i \neq 0\}}(\bar{w}), \end{aligned}$$

where the second inequality uses the lower semicontinuity of each  $p_i \mathbb{1}_{\{\bar{w}^i \neq 0\}}(\bar{w})$ . The lower semicontinuity of  $\sum_{i=1}^n p_i \mathbb{1}_{\{w^i \neq 0\}}(w)$  proves that its lower level set  $C$  is closed.  $\blacksquare$

## Appendix B. Section 5 Proofs

**Proof of Proposition 2** Given that  $\mathbb{E}[L_0(\xi)^2] < \infty$  it holds that  $L_0(\xi)$  is integrable,  $\mathbb{E}[L_0(\xi)] < \infty$  (Folland, 1999, Proposition 6.12). For any  $w, w' \in \bar{B}_\kappa$ ,

$$\begin{aligned} |f(w) - f(w')| &= |\mathbb{E}[F(w, \xi) - F(w', \xi)]| \\ &\leq \mathbb{E}[|F(w, \xi) - F(w', \xi)|] \\ &\leq \mathbb{E}[L_0(\xi) \|w - w'\|_2] \\ &= L_0 \|w - w'\|_2, \end{aligned}$$

where the first inequality uses Jensen's inequality and the second inequality holds given that (4) holds for almost all  $\xi \in \mathbb{R}^p$ .  $\blacksquare$

**Proof of Proposition 3** Let  $\bar{w} \in \bar{B}_\beta$ , let  $\{h_j\} \subset \mathbb{R}$  be a sequence such that  $0 < |h_j| \leq \kappa - (\beta + \frac{\alpha}{2})$  with  $\lim_{j \rightarrow \infty} h_j \rightarrow 0$ , and let  $e_i$  equal the  $i^{\text{th}}$  unit coordinate vector. Given the Lipschitz continuity of  $f(w)$  over  $\bar{B}_\kappa$ ,

$$\lim_{j \rightarrow \infty} h_j^{-1} (f(\bar{w} + u + h_j e_i) - f(\bar{w} + u)) = \tilde{\nabla}_i f(\bar{w} + u)$$

for almost every  $u \in \bar{B}_{\alpha/2}$ . For all  $j \in \mathbb{N}$ ,

$$|h_j^{-1} |f(\bar{w} + u + h_j e_i) - f(\bar{w} + u)| \leq L_0 |h_j^{-1}| |h_j e_i| = L_0 \in L^1(P_u).$$

Applying the dominated convergence theorem,

$$\begin{aligned} \mathbb{E} \left[ \tilde{\nabla}_i f(\bar{w} + u) \right] &= \lim_{j \rightarrow \infty} \mathbb{E} [h_j^{-1} (f(\bar{w} + u + h_j e_i) - f(\bar{w} + u))] \\ &= \lim_{j \rightarrow \infty} h_j^{-1} (f_\alpha(\bar{w} + h_j e_i) - f_\alpha(\bar{w})). \end{aligned}$$

Given that  $\mathbb{E}[|\tilde{\nabla}_i f(\bar{w} + u)|] \leq L_0$  (Metel and Takeda, 2022, Lemma 6), we can use Fubini's theorem,

$$\begin{aligned} \mathbb{E}\left[\tilde{\nabla}_i f(\bar{w} + u)\right] &= \frac{1}{\alpha^d} \int_{-\alpha/2}^{\alpha/2} \cdots \int_{-\alpha/2}^{\alpha/2} \int_{-\alpha/2}^{\alpha/2} \cdots \int_{-\alpha/2}^{\alpha/2} \int_{-\alpha/2}^{\alpha/2} \tilde{\nabla}_i f(\bar{w} + u) du_i du_1 \cdots du_{i-1} du_{i+1} \cdots du_d \\ &= \frac{1}{\alpha^d} \int_{-\alpha/2}^{\alpha/2} \cdots \int_{-\alpha/2}^{\alpha/2} \int_{-\alpha/2}^{\alpha/2} \cdots \int_{-\alpha/2}^{\alpha/2} df_i(\bar{w}, u_{\setminus i}) du_1 \cdots du_{i-1} du_{i+1} \cdots du_d \\ &= \alpha^{-1} \mathbb{E}[df_i(\bar{w}, u_{\setminus i})], \end{aligned}$$

where the second equality uses the fundamental theorem of calculus for Lebesgue integrals given that  $\tilde{\nabla}_i f(\bar{w} + u) = \nabla_i f(\bar{w} + u)$  for almost all  $u \in [-\alpha/2, \alpha/2]^d$ : For a fixed  $u'_{\setminus i} \in [-\alpha/2, \alpha/2]^{d-1}$ , let  $u'(u_i) := [u'_1, \dots, u'_{i-1}, u_i, u'_{i+1}, \dots, u'_d]^T$  for  $u_i \in [-\alpha/2, \alpha/2]$ . Assume that  $u'_{\setminus i}$  is chosen such that  $f(\bar{w} + u'(u_i))$  is differentiable with  $\tilde{\nabla}_i f(\bar{w} + u'(u_i)) = \nabla_i f(\bar{w} + u'(u_i))$  for almost all  $u_i \in [-\alpha/2, \alpha/2]$ . Given that  $f(\bar{w} + u'(u_i))$  is absolutely continuous for  $u_i \in [-\alpha/2, \alpha/2]$ ,  $\int_{-\alpha/2}^{\alpha/2} \tilde{\nabla}_i f(\bar{w} + u'(u_i)) du_i = df_i(\bar{w}, u'_{\setminus i})$ . Since this holds for almost all  $u'_{\setminus i} \in [-\alpha/2, \alpha/2]^{d-1}$ , the second equality holds.

The equality  $\mathbb{E}[\tilde{\nabla} f(\bar{w} + u)] = \mathbb{E}[\tilde{\nabla} F(\bar{w} + u, \xi)]$  holds by (Metel and Takeda, 2022, Property 2), where it is shown that for almost all  $u \in \bar{B}_{\alpha/2}$ ,  $\tilde{\nabla} f(\bar{w} + u) = \mathbb{E}_{\xi}[\tilde{\nabla} F(\bar{w} + u, \xi)]$  in our problem setting.

As  $\bar{w} \in \bar{B}_{\beta}$  and the sequence  $\{h_i\} \subset \bar{B}_{\kappa - (\beta + \frac{\alpha}{2})}$  were arbitrary, for all  $w \in \bar{B}_{\beta}$ ,

$$\begin{aligned} \alpha^{-1} \mathbb{E}[df_i(w, u_{\setminus i})] &= \mathbb{E}[\tilde{\nabla} F_i(w + u, \xi)] \\ &= \mathbb{E}[\tilde{\nabla}_i f(w + u)] \\ &= \lim_{h \rightarrow 0} h^{-1} (f_{\alpha}(w + he_i) - f_{\alpha}(w)) \\ &= \frac{\partial f_{\alpha}}{\partial w_i}(w) \\ &= \nabla_i f_{\alpha}(w), \end{aligned} \tag{13}$$

where the third equality holds using the sequential criterion of a limit, and the last equality holds given that the partial derivatives are (Lipschitz) continuous: For any  $w, w' \in \bar{B}_{\beta}$ ,

$$\begin{aligned} &|\alpha^{-1} \mathbb{E}[df_i(w, u_{\setminus i})] - \alpha^{-1} \mathbb{E}[df_i(w', u_{\setminus i})]| \\ &\leq \alpha^{-1} \mathbb{E}[|df_i(w, u_{\setminus i}) - df_i(w', u_{\setminus i})|] \\ &\leq 2\alpha^{-1} L_0 \|w - w'\|_2, \end{aligned}$$

where a proof of the last inequality can be found at (15). Given that for any  $w \in \bar{B}_{\beta}$ ,

$$\alpha^{-1} \mathbb{E}[|dF_i(w, u_{\setminus i}, \xi)|] \leq \alpha^{-1} \mathbb{E}[L_0(\xi)\alpha] = L_0,$$

Fubini's theorem can be applied:

$$\begin{aligned} \alpha^{-1} \mathbb{E}[dF_i(w, u_{\setminus i}, \xi)] &= \alpha^{-1} \mathbb{E}_{u_{\setminus i}}[\mathbb{E}_{\xi}[dF_i(w, u_{\setminus i}, \xi)]] \\ &= \alpha^{-1} \mathbb{E}[df_i(w, u_{\setminus i})] \\ &= \nabla_i f_{\alpha}(w) \end{aligned}$$

from (13). ■

**Proof of Proposition 4** Given that  $u \in \overline{B}_{\frac{\alpha}{2}}$ ,  $w + u \in \overline{B}(w, \sqrt{d}\frac{\alpha}{2}) \subset \overline{B}_\kappa$ , hence  $\overline{\partial}_{\hat{\alpha}}f(w) = \text{co}\{\overline{\partial}f(x) : x \in \overline{B}(w, \hat{\alpha})\}$  is well defined. The gradient  $\nabla f(w + u) \in \overline{\partial}f(w + u)$  wherever  $f(w + u)$  is differentiable (Clarke, 1990, Proposition 2.2.2), which is for almost all  $u \in \overline{B}_{\frac{\alpha}{2}}$ . It follows that for almost all  $u \in \overline{B}_{\frac{\alpha}{2}}$ ,  $\tilde{\nabla}f(w + u) \in \overline{\partial}_{\hat{\alpha}}f(w)$  and  $\nabla f_\alpha(w) = \mathbb{E}[\tilde{\nabla}f(w + u)] \in \overline{\partial}_{\hat{\alpha}}f(w)$ . ■

**Proof of Proposition 5**

1. For all  $w, w' \in \overline{B}_\beta$ ,

$$|f_\alpha(w) - f_\alpha(w')| = |\mathbb{E}[f(w + u) - f(w' + u)]| \leq L_0 \|w - w'\|_2.$$

2. For all  $w \in \overline{B}_\beta$ ,

$$|f_\alpha(w) - f(w)| = |\mathbb{E}[f(w + u) - f(w)]| \leq L_0 \mathbb{E}[\|u\|_2] \leq L_0 \sqrt{\frac{d}{12}} a,$$

where the expected distance from  $u$  to the origin is bounded by

$$\mathbb{E}[\|u\|_2] \leq \sqrt{\mathbb{E}[\sum_{i=1}^d u_i^2]} = \sqrt{d} \sqrt{\mathbb{E}[u_i^2]} = \sqrt{d} \frac{\alpha}{\sqrt{12}}.$$

3. This can be proven by contradiction. Assuming that  $f_\alpha(w_\alpha^*) - f(w^*) > \alpha L_0 \sqrt{\frac{d}{12}}$  and given that  $f_\alpha(w^*) - f(w^*) \leq \alpha L_0 \sqrt{\frac{d}{12}}$  from statement 2,

$$f_\alpha(w_\alpha^*) > \alpha L_0 \sqrt{\frac{d}{12}} + f(w^*) \geq f_\alpha(w^*),$$

contradicting the optimality of  $w_\alpha^*$  for  $f_\alpha(w)$ . Similarly if  $f(w^*) - f_\alpha(w_\alpha^*) > \alpha L_0 \sqrt{\frac{d}{12}}$ , statement 2 gives  $f(w_\alpha^*) - f_\alpha(w_\alpha^*) \leq \alpha L_0 \sqrt{\frac{d}{12}}$ , from which

$$f(w^*) > \alpha L_0 \sqrt{\frac{d}{12}} + f_\alpha(w_\alpha^*) \geq f_\alpha(w^*),$$

contradicting the optimality of  $w^*$  for  $f(w)$ .

4. For any two values  $w, w' \in \overline{B}_\beta$ ,

$$\begin{aligned} |f_\alpha(w) - f_\alpha(w')| &= |\mathbb{E}[f(w + u)] - \mathbb{E}[f(w' + u)]| \\ &= |\mathbb{E}[f(w + u) - f(w' + u)]| \\ &\leq \mathbb{E}[L_0 \|w - w'\|_2] \\ &\leq L_0 \sqrt{\sum_{i=1}^d (2\beta)^2} \\ &= L_0 \sqrt{d} 2\beta. \end{aligned}$$

■

**Proof of Proposition 6** From Proposition 3,

$$\begin{aligned} \|\nabla f_\alpha(w) - \nabla f_\alpha(w')\|_2 &= \alpha^{-1} \|\mathbb{E}[df(w, u) - df(w', u)]\|_2 \\ &= \alpha^{-1} \sqrt{\sum_{i=1}^d (\mathbb{E}[df_i(w, u_{\setminus i}) - df_i(w', u_{\setminus i})])^2}. \end{aligned} \quad (14)$$

Focusing on  $df_i(w, u_{\setminus i}) - df_i(w', u_{\setminus i})$ ,

$$\begin{aligned} &|df_i(w, u_{\setminus i}) - df_i(w', u_{\setminus i})| \\ &= |f(w_1 + u_1, \dots, w_{i-1} + u_{i-1}, w_i + \frac{\alpha}{2}, w_{i+1} + u_{i+1}, \dots, w_d + u_d) \\ &\quad - f(w_1 + u_1, \dots, w_{i-1} + u_{i-1}, w_i - \frac{\alpha}{2}, w_{i+1} + u_{i+1}, \dots, w_d + u_d) \\ &\quad - f(w'_1 + u_1, \dots, w'_{i-1} + u_{i-1}, w'_i + \frac{\alpha}{2}, w'_{i+1} + u_{i+1}, \dots, w'_d + u_d) \\ &\quad + f(w'_1 + u_1, \dots, w'_{i-1} + u_{i-1}, w'_i - \frac{\alpha}{2}, w'_{i+1} + u_{i+1}, \dots, w'_d + u_d)| \\ &\leq |f(w_1 + u_1, \dots, w_{i-1} + u_{i-1}, w_i + \frac{\alpha}{2}, w_{i+1} + u_{i+1}, \dots, w_d + u_d) \\ &\quad - f(w'_1 + u_1, \dots, w'_{i-1} + u_{i-1}, w'_i + \frac{\alpha}{2}, w'_{i+1} + u_{i+1}, \dots, w'_d + u_d)| \\ &\quad + |f(w'_1 + u_1, \dots, w'_{i-1} + u_{i-1}, w'_i - \frac{\alpha}{2}, w'_{i+1} + u_{i+1}, \dots, w'_d + u_d) \\ &\quad - f(w_1 + u_1, \dots, w_{i-1} + u_{i-1}, w_i - \frac{\alpha}{2}, w_{i+1} + u_{i+1}, \dots, w_d + u_d)| \\ &\leq 2L_0 \|w - w'\|_2. \end{aligned} \quad (15)$$

Plugging (15) into (14),

$$\begin{aligned} \|\nabla f_\alpha(w) - \nabla f_\alpha(w')\|_2 &\leq \alpha^{-1} \sqrt{\sum_{i=1}^d (2L_0 \|w - w'\|_2)^2} \\ &= \alpha^{-1} \sqrt{d} 2L_0 \|w - w'\|_2. \end{aligned}$$

■

**Proof of Proposition 7** To streamline the proof, let  $G := \frac{1}{M} \sum_{i=1}^M g^i$ , where

$$g^i := \begin{cases} \alpha^{-1} dF(w, u^i, \xi^i) & \text{for 1 and 3} \\ \tilde{\nabla} F(w + u^i, \xi^i) & \text{for 2 and 4.} \end{cases}$$

Proof of 1 and 2:

$$\begin{aligned}
\mathbb{E}[\|\nabla f_\alpha(w) - G\|_2^2] &= \mathbb{E}[\|\mathbb{E}[G] - G\|_2^2] \\
&= \mathbb{E}\left[\sum_{j=1}^d (\mathbb{E}[G_j] - G_j)^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^d \left(\frac{1}{M} \sum_{i=1}^M (\mathbb{E}[G_j] - g_j^i)\right)^2\right] \\
&= \frac{1}{M^2} \mathbb{E}\left[\sum_{j=1}^d \left(\sum_{i=1}^M (\mathbb{E}[G_j] - g_j^i)^2 + 2 \sum_{i=1}^M \sum_{l=1}^{i-1} (\mathbb{E}[G_j] - g_j^i)(\mathbb{E}[G_j] - g_j^l)\right)\right] \\
&= \frac{1}{M^2} \sum_{j=1}^d \sum_{i=1}^M \mathbb{E}[(\mathbb{E}[G_j] - g_j^i)^2] \\
&= \frac{1}{M} \sum_{j=1}^d \mathbb{E}[(\mathbb{E}[G_j] - g_j^i)^2] \\
&= \frac{1}{M} \sum_{j=1}^d (\mathbb{E}[G_j]^2 - 2\mathbb{E}[G_j]\mathbb{E}[g_j^i] + \mathbb{E}[(g_j^i)^2]) \\
&= \frac{1}{M} \sum_{j=1}^d (-\mathbb{E}[G_j]^2 + \mathbb{E}[(g_j^i)^2]) \\
&\leq \frac{1}{M} \sum_{j=1}^d \mathbb{E}[(g_j^i)^2].
\end{aligned}$$

For 1:

$$\begin{aligned}
\frac{1}{M} \sum_{j=1}^d \mathbb{E}[(g_j^i)^2] &= \frac{1}{M} \sum_{j=1}^d \mathbb{E}[(\alpha^{-1} dF_j(w, u_{\setminus j}^i, \xi^i))^2] \\
&= \frac{1}{\alpha^2 M} \sum_{j=1}^d \mathbb{E}[dF_j(w, u_{\setminus j}^i, \xi^i)^2] \\
&\leq \frac{1}{\alpha^2 M} \sum_{j=1}^d \mathbb{E}[L_0(\xi)^2 \alpha^2] \\
&= \frac{d}{M} \mathbb{E}[L_0(\xi)^2]. \tag{16}
\end{aligned}$$

For 2:

$$\begin{aligned}
 \frac{1}{M} \sum_{j=1}^d \mathbb{E}[(g_j^i)^2] &= \frac{1}{M} \sum_{j=1}^d \mathbb{E}[\tilde{\nabla} F(w + u^i, \xi^i)^2] \\
 &= \frac{1}{M} \mathbb{E}[\|\tilde{\nabla} F(w + u^i, \xi^i)\|_2^2] \\
 &\leq \frac{1}{M} \mathbb{E}[L_0(\xi)^2],
 \end{aligned} \tag{17}$$

where the last inequality follows from (Metel and Takeda, 2022, Lemma 6).

Proof of 3 and 4:

$$\begin{aligned}
 \mathbb{E}[\|G\|_2^2] &= \mathbb{E}[\|\frac{1}{M} \sum_{i=1}^M g^i\|_2^2] \\
 &= \mathbb{E}[\sum_{j=1}^d (\frac{1}{M} \sum_{i=1}^M g_j^i)^2] \\
 &\leq \mathbb{E}[\sum_{j=1}^d \frac{1}{M} \sum_{i=1}^M (g_j^i)^2] \\
 &= \mathbb{E}[\sum_{j=1}^d (g_j^i)^2],
 \end{aligned}$$

where the inequality uses Jensen's inequality. For 3, from (16):

$$\mathbb{E}[\sum_{j=1}^d (g_j^i)^2] \leq d \mathbb{E}[L_0(\xi^i)^2],$$

and for 4, from (17):

$$\mathbb{E}[\sum_{j=1}^d (g_j^i)^2] \leq \mathbb{E}[L_0(\xi^i)^2].$$

■

## Appendix C. Section 6 Proofs

**Proof of Proposition 8** Following (Rockafellar and Wets, 2009, Chapter 1.G.), the projection of a point  $w$  onto  $C \cap \bar{B}_\beta$  can be written as

$$\Pi_{C \cap \bar{B}_\beta}(w) = \operatorname{argmin}_{x \in \mathbb{R}^d} \left\{ \delta_{C \cap \bar{B}_\beta}(x) + \sum_{i=1}^n \|x^i - w^i\|_2^2 \right\}.$$

For the squared distance function of a point  $w \in \mathbb{R}^d$  from  $C \cap \overline{B}_\beta$ ,

$$d_{C \cap \overline{B}_\beta}^2(w) = \inf_{x \in \mathbb{R}^d} \left\{ \delta_{C \cap \overline{B}_\beta}(x) + \sum_{i=1}^n \|x^i - w^i\|_2^2 \right\},$$

let  $x^*$  be a minimizer. If  $(x^*)^i \neq 0$  then it will be the optimal solution of

$$\begin{aligned} \min_{x \in \mathbb{R}^{d_i}} \quad & \|x^i - w^i\|_2^2 \\ \text{s.t.} \quad & |x_j^i| \leq \beta \quad j = 1, 2, \dots, d_i, \end{aligned}$$

equal to  $(x^*)^i_j = \text{sgn}(w_j^i) \min(|w_j^i|, \beta)$  for  $j = 1, 2, \dots, d_i$ , with the optimal objective value of  $\|\max(|w^i| - \beta, 0)\|_2^2$ . The function  $d_{C \cap \overline{B}_\beta}^2$  can then be written as

$$\begin{aligned} d_{C \cap \overline{B}_\beta}^2(w) = \min_{z \in \{0,1\}^n} \quad & \sum_{i=1}^n z_i \|\max(|w^i| - \beta, 0)\|_2^2 + (1 - z_i) \|w^i\|_2^2 \\ \text{s.t.} \quad & \sum_{i=1}^n z_i p_i \leq m, \end{aligned}$$

where each  $z_i$  decides if  $x^i$  is allowed to be non-zero. This optimization problem has the same set of optimal solutions as

$$\begin{aligned} \min_{z \in \{0,1\}^n} \quad & \sum_{i=1}^n z_i (\|\max(|w^i| - \beta, 0)\|_2^2 - \|w^i\|_2^2) \\ \text{s.t.} \quad & \sum_{i=1}^n z_i p_i \leq m, \end{aligned}$$

and

$$\begin{aligned} \max_{z \in \{0,1\}^n} \quad & \sum_{i=1}^n z_i (\|w^i\|_2^2 - \|\max(|w^i| - \beta, 0)\|_2^2) \\ \text{s.t.} \quad & \sum_{i=1}^n z_i p_i \leq m, \end{aligned}$$

written in the form of a maximization to match the standard format of knapsack problems. ■

**Proof of Proposition 10** Let  $RHS$  equal the right-hand side of (7) and for simplicity let  $I := I(w)$  and  $J := J(w)$ . If  $w = 0$ , then  $RHS = 0 \in \widehat{N}(w, C \cap \overline{B}_\beta)$ . Assuming  $w \neq 0$ , let  $v \in RHS$  and  $\hat{m} := \min_{i \in I} \|w^i\|_2/2$ . For all  $x \in B(w, \hat{m}) \cap C \cap \overline{B}_\beta =: D$ ,  $I \subseteq I(x)$ . Given that  $x \in C$  and  $I \subseteq I(x)$ ,  $I(x) \subseteq I \cup J$ : If not, then there exists a  $j \in I(x) \setminus I$  such that  $j \notin J$ , but this implies that  $\sum_{i \in I(x)} p_i + p_j \geq \sum_{i \in I} p_i + p_j > m$ , which contradicts that  $x \in C$ . Given that  $I(x) \subseteq I \cup J$  for all  $x \in D$ ,

$$\langle v, x - w \rangle = \sum_{i \notin I \cup J} \langle v^i, x^i - w^i \rangle + \sum_{i \in I \cup J} \sum_{j \in [d_i]} v_j^i (x_j^i - w_j^i) \leq 0 \quad (18)$$



since for  $i \notin I \cup J$ ,  $x^i = w^i = 0$  and for  $i \in I \cup J$ , given that  $x \in \overline{B}_\beta$ ,  $v_j^i(x_j^i - w_j^i) \leq 0$  from (7). From (18),

$$\inf_{\gamma > 0} \left( \sup_{\substack{0 < \|x-w\|_2 < \gamma \\ x \in C \cap \overline{B}_\beta}} \frac{\langle v, x-w \rangle}{\|x-w\|_2} \right) \leq \sup_{\substack{0 < \|x-w\|_2 < \hat{m} \\ x \in C \cap \overline{B}_\beta}} \frac{\langle v, x-w \rangle}{\|x-w\|_2} \leq 0$$

and  $v \in \widehat{N}(w, C \cap \overline{B}_\beta)$ .

For a  $v \in \widehat{N}(w, C \cap \overline{B}_\beta)$ , assume there exists an  $i \in I \cup J$  and  $j \in [d_i]$  for which it does not hold that

$$v_j^i \in \begin{cases} \mathbb{R}_{\geq 0} & \text{if } w_j^i = \beta \\ 0 & \text{if } |w_j^i| < \beta \\ \mathbb{R}_{\leq 0} & \text{if } w_j^i = -\beta. \end{cases}$$

Consider the sequence  $\{x_k\}$  with elements equal to

$$(x_m^l)_k = \begin{cases} w_m^l + \epsilon_k v_m^l & \text{if } l = i \text{ and } m = j \\ w_m^l & \text{otherwise} \end{cases}$$

with

$$\begin{cases} 0 < \epsilon_k \leq -2\frac{\beta}{v_j^i} & \text{if } v_j^i < 0 \text{ and } w_j^i = \beta \\ 0 < \epsilon_k \leq \frac{\beta - |w_j^i|}{|v_j^i|} & \text{if } v_j^i \neq 0 \text{ and } |w_j^i| < \beta \\ 0 < \epsilon_k \leq 2\frac{\beta}{v_j^i} & \text{if } v_j^i > 0 \text{ and } w_j^i = -\beta, \end{cases}$$

to ensure that  $\{x_k\} \subset \overline{B}_\beta$ , and assume that  $\epsilon_k \rightarrow 0$ . It holds that  $x_k \xrightarrow{C \cap \overline{B}_\beta} w$ , since  $I(x_k) \subseteq I \cup i$  for all  $k \in \mathbb{N}$ . For any  $\gamma > 0$ , there exists a  $K_\gamma \in \mathbb{N}$  such that for  $k > K_\gamma$ ,  $0 < \|x_k - w\|_2 < \gamma$ , hence

$$\sup_{\substack{0 < \|x-w\|_2 < \gamma \\ x \in C \cap \overline{B}_\beta}} \frac{\langle v, x-w \rangle}{\|x-w\|_2} \geq \frac{\langle v, x_k - w \rangle}{\|x_k - w\|_2} = \frac{\epsilon_k (v_j^i)^2}{\epsilon_k |v_j^i|} = |v_j^i| > 0,$$

which contradicts that  $v \in \widehat{N}(w, C \cap \overline{B}_\beta)$ . ■

**Proof of Proposition 12** Let *RHS* equal the right-hand side of (8) and let  $I := I(w)$  and  $J := J(w)$ . For any  $v \in N(w, C \cap \overline{B}_\beta)$ , there exists sequences  $x_k \xrightarrow{C \cap \overline{B}_\beta} w$  and  $v_k \rightarrow v$  with  $v_k \in \widehat{N}(x_k, C \cap \overline{B}_\beta)$  for all  $k \in \mathbb{N}$ . We first want to show that for  $k \in \mathbb{N}$  sufficiently large, there exist  $X(v_k) \in Y$  such that  $X(v_k) \subseteq I(x_k) \cup J(x_k)$ , implying that  $v_k \in \text{RHS}$ .

Assuming  $w \neq 0$ , there exists an  $N \in \mathbb{N}$  such that for all  $k > N$ ,  $\max_{i \in [n]} \|x_k^i - w^i\|_2 < \min_{i \in I} \|w^i\|_2 / 2$ , implying that  $I \subseteq I(x_k)$ . Given that  $I \subseteq I(x_k)$ , there exists an  $X(v_k) \in Y$

such that  $X(v_k) \subseteq I(x_k) \cup J(x_k)$ : We can choose  $X(v_k)$  such that  $I(x_k) \subseteq X(v_k)$  given that  $I \subseteq I(x_k)$  and  $x_k \in C$ . If  $X(v_k) \not\subseteq I(x_k) \cup J(x_k)$ , then there exists a  $j \in X(v_k) \setminus I(x_k)$  such that  $j \notin J(x_k)$ , but this implies that  $\sum_{i \in X(v_k)} p_i \geq \sum_{i \in I(x_k)} p_i + p_j > m$ , which contradicts that  $X(v_k) \in Y$ .

For the case when  $w = 0$ ,  $I = \{\emptyset\}$ , hence we can choose  $X(v_k)$  such that  $I(x_k) \subseteq X(v_k)$ , from which it must hold again that  $X(v_k) \subseteq I(x_k) \cup J(x_k)$ . Given the existence of an  $X(v_k) \in Y$  such that  $X(v_k) \subseteq I(x_k) \cup J(x_k)$ ,  $v_k \in RHS$  for all  $k > N$ , from which it follows that  $v \in RHS$  given that RHS is a closed set: For any  $w \in \mathbb{R}^d$ ,

$$N_j^i(w) := \left\{ y \in \mathbb{R}^d : y_j^i \in \begin{cases} \mathbb{R}_{\geq 0} & \text{if } w_j^i = \beta \\ 0 & \text{if } |w_j^i| < \beta \\ \mathbb{R}_{\leq 0} & \text{if } w_j^i = -\beta \end{cases} \right\}$$

is a closed set, the intersection of closed sets

$$\bigcap_{\substack{\forall i \in X \\ \forall j \in [d_i]}} N_j^i(w) \tag{19}$$

for an  $X \in Y$  is closed, and the finite union of sets (19) over  $X \in Y$

$$RHS = \bigcup_{X \in Y} \bigcap_{\substack{\forall i \in X \\ \forall j \in [d_i]}} N_j^i(w)$$

is closed.

Let  $v \in RHS$  and  $X(v) \in Y$  such that  $\forall i \in X(v)$  and  $\forall j \in [d_i]$ ,

$$v_j^i \in \begin{cases} \mathbb{R}_{\geq 0} & \text{if } w_j^i = \beta \\ 0 & \text{if } |w_j^i| < \beta \\ \mathbb{R}_{\leq 0} & \text{if } w_j^i = -\beta. \end{cases}$$

Consider the sequence  $\{x_k\}$  equal to

$$x_k^i = \begin{cases} w^i & \text{if } i \in I \\ \epsilon_k^i & \text{if } i \in X(v) \setminus I \\ 0 & \text{otherwise,} \end{cases}$$

where for a  $j \in [d_i]$ ,  $0 < (\epsilon_j^i)_k \leq \beta$  and  $(\epsilon_j^i)_k \rightarrow 0$ , and  $(\epsilon_l^i)_k = 0$  for all  $l \in [d_i] \setminus \{j\}$  and  $k \in \mathbb{N}$ . It holds that  $x_k \xrightarrow{C \cap \bar{B}_\beta} w$ , the sets  $I(x_k) = X(v)$  and  $J(x_k) = \{\emptyset\}$ , hence  $v \in \widehat{N}(x_k, C)$  for all  $k \in \mathbb{N}$ , proving that  $v \in N(w, C)$ . We note that for the case  $X(v) = I$ ,  $v \in \widehat{N}(w, C) \subseteq N(w, C)$  (Mordukhovich, 2013, Eq. (1.6))  $\blacksquare$

## Appendix D. Section 7 Proofs

**Proof of Proposition 14** The measurability of  $\Pi_{S \cap \overline{B}_\beta}(\cdot)$  follows from (Rockafellar and Wets, 2009, Exercise 14.17 (b)). In particular, the set  $S \cap \overline{B}_\beta$  is closed and can be considered as a constant (measurable) set valued function  $S \cap \overline{B}_\beta : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ . Given that the projection operator is a closed (in particular a nonempty and compact) set-valued mapping (Rockafellar and Wets, 2009, Exercise 1.20), there exists a measurable selection (Rockafellar and Wets, 2009, Corollary 14.6).

The restriction that  $\epsilon \leq \frac{\alpha}{2}$  is to ensure that  $\overline{\partial}_\epsilon f(w)$  is well defined given that it is only assumed that  $\kappa > \beta + \frac{\alpha}{2}$ . The set valued mapping  $\overline{\partial}_\epsilon f(w)$  is outer semicontinuous for  $w \in \overline{B}_\beta$ , see (Goldstein, 1977, Lemma 2.6) for a proof that (Rockafellar and Wets, 2009, Definition 5.4) holds, as is  $N(w, S \cap \overline{B}_\beta)$  for  $w \in \mathbb{R}^d$  given that  $S \cap \overline{B}_\beta$  is closed (Rockafellar and Wets, 2009, Theorem 5.7 (a); Page 202). Since  $\overline{\partial}_\epsilon f(w)$  is bounded, given that  $\overline{\partial}_\epsilon f(w) \subseteq \overline{B}(0, L_0)$  (Clarke, 1990, Proposition 2.1.2 (a)),  $\overline{\partial}_\epsilon f(w) + N(w, S \cap \overline{B}_\beta)$  is outer semicontinuous for  $w \in \overline{B}_\beta$  (Rockafellar and Wets, 2009, Proposition 5.51 (b)). The function  $\text{dist}(0, \overline{\partial}_\epsilon f(w) + N(w, S \cap \overline{B}_\beta))$  is then lower semicontinuous for  $w \in \overline{B}_\beta$  (Rockafellar and Wets, 2009, Proposition 5.11), hence Borel measurable. The Borel measurability of  $\text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \overline{B}_\beta))$  holds since  $f_\alpha(w)$  is Lipschitz continuous (Proposition 5.1), continuously differentiable (Proposition 6), hence  $\nabla f_\alpha(w) = \overline{\partial} f_\alpha(w)$  (Clarke, 1990, Page 10).  $\blacksquare$

**Proof of Theorem 15** For simplicity let

$$G^k := \begin{cases} \frac{1}{M\alpha} \sum_{i=1}^M dF(w^k, u^{k,i}, \xi^{k,i}) & \text{if using (1)} \\ \frac{1}{M} \sum_{i=1}^M \tilde{\nabla} F(w^k + u^{k,i}, \xi^{k,i}) & \text{if using (2)} \end{cases}$$

in SPA, where we assume  $k \in [K]$ . Given that

$$w^{k+1} \in \underset{x \in \mathbb{R}^d}{\text{argmin}} \{ \delta_{S \cap \overline{B}_\beta}(x) + \|x - (w^k - \eta G^k)\|_2^2 \} \quad (20)$$

and  $\|x - (w^k - \eta G^k)\|_2^2$  is differentiable (Rockafellar and Wets, 2009, Theorem 6.12),

$$\begin{aligned} & -2\eta G^k - 2(w^{k+1} - w^k) \in N(w^{k+1}, S \cap \overline{B}_\beta) \\ \implies & -G^k - \eta^{-1}(w^{k+1} - w^k) \in N(w^{k+1}, S \cap \overline{B}_\beta) \\ \implies & \nabla f_\alpha(w^{k+1}) - G^k - \eta^{-1}(w^{k+1} - w^k) \in \nabla f_\alpha(w^{k+1}) + N(w^{k+1}, S \cap \overline{B}_\beta), \end{aligned}$$

where the second inclusion holds since  $N(w^{k+1}, S \cap \overline{B}_\beta)$  is a cone. The final inclusion implies that

$$\begin{aligned}
 & \text{dist}(0, \nabla f_\alpha(w^{k+1}) + N(w^{k+1}, S \cap \overline{B}_\beta))^2 \\
 & \leq \|\nabla f_\alpha(w^{k+1}) - G^k - \eta^{-1}(w^{k+1} - w^k)\|_2^2 \\
 & = \|\nabla f_\alpha(w^{k+1}) - G^k\|_2^2 - 2\eta^{-1}\langle \nabla f_\alpha(w^{k+1}) - G^k, w^{k+1} - w^k \rangle + \eta^{-2}\|w^{k+1} - w^k\|_2^2 \\
 & = \|\nabla f_\alpha(w^{k+1}) - \nabla f_\alpha(w^k) + \nabla f_\alpha(w^k) - G^k\|_2^2 \\
 & \quad - 2\eta^{-1}\langle \nabla f_\alpha(w^{k+1}) - G^k, w^{k+1} - w^k \rangle + \eta^{-2}\|w^{k+1} - w^k\|_2^2 \\
 & \leq \frac{3}{2}\|\nabla f_\alpha(w^{k+1}) - \nabla f_\alpha(w^k)\|_2^2 + 3\|\nabla f_\alpha(w^k) - G^k\|_2^2 \\
 & \quad - 2\eta^{-1}\langle \nabla f_\alpha(w^{k+1}) - G^k, w^{k+1} - w^k \rangle + \eta^{-2}\|w^{k+1} - w^k\|_2^2 \\
 & \leq 6\alpha^{-2}dL_0^2\|w^{k+1} - w^k\|_2^2 + 3\|\nabla f_\alpha(w^k) - G^k\|_2^2 \\
 & \quad - 2\eta^{-1}\langle \nabla f_\alpha(w^{k+1}) - G^k, w^{k+1} - w^k \rangle + \eta^{-2}\|w^{k+1} - w^k\|_2^2 \\
 & = (6\alpha^{-2}dL_0^2 + \eta^{-2})\|w^{k+1} - w^k\|_2^2 + 3\|\nabla f_\alpha(w^k) - G^k\|_2^2 \\
 & \quad - 2\eta^{-1}\langle \nabla f_\alpha(w^{k+1}) - \nabla f_\alpha(w^k) + \nabla f_\alpha(w^k) - G^k, w^{k+1} - w^k \rangle \\
 & \leq (6\alpha^{-2}dL_0^2 + \eta^{-2})\|w^{k+1} - w^k\|_2^2 + 3\|\nabla f_\alpha(w^k) - G^k\|_2^2 \\
 & \quad + 4\alpha^{-1}\sqrt{d}L_0\eta^{-1}\|w^{k+1} - w^k\|_2^2 - 2\eta^{-1}\langle \nabla f_\alpha(w^k) - G^k, w^{k+1} - w^k \rangle \\
 & = (6\alpha^{-2}dL_0^2 + 4\alpha^{-1}\sqrt{d}L_0\eta^{-1} + \eta^{-2})\|w^{k+1} - w^k\|_2^2 + 3\|\nabla f_\alpha(w^k) - G^k\|_2^2 \\
 & \quad - 2\eta^{-1}\langle \nabla f_\alpha(w^k) - G^k, w^{k+1} - w^k \rangle, \tag{21}
 \end{aligned}$$

where the second inequality uses Young's inequality:

$$\begin{aligned}
 & \|\nabla f_\alpha(w^{k+1}) - \nabla f_\alpha(w^k) + \nabla f_\alpha(w^k) - G^k\|_2^2 \\
 & = \|\nabla f_\alpha(w^{k+1}) - \nabla f_\alpha(w^k)\|_2^2 + 2\langle \frac{1}{\sqrt{2}}(\nabla f_\alpha(w^{k+1}) - \nabla f_\alpha(w^k)), \sqrt{2}(\nabla f_\alpha(w^k) - G^k) \rangle \\
 & \quad + \|\nabla f_\alpha(w^k) - G^k\|_2^2 \\
 & \leq \frac{3}{2}\|\nabla f_\alpha(w^{k+1}) - \nabla f_\alpha(w^k)\|_2^2 + 3\|\nabla f_\alpha(w^k) - G^k\|_2^2,
 \end{aligned}$$

and the third and fourth inequalities use the smoothness of  $f_\alpha(w)$ . By the optimality of  $w^{k+1}$  in (20),

$$\begin{aligned}
 & \delta_{S \cap \overline{B}_\beta}(w^{k+1}) + \|w^{k+1} - (w^k - \eta G^k)\|_2^2 \leq \delta_{S \cap \overline{B}_\beta}(w^k) + \|\eta G^k\|_2^2 \\
 & \implies \delta_{S \cap \overline{B}_\beta}(w^{k+1}) + \|w^{k+1} - w^k\|_2^2 + 2\eta\langle w^{k+1} - w^k, G^k \rangle \leq \delta_{S \cap \overline{B}_\beta}(w^k) \\
 & \implies \|w^{k+1} - w^k\|_2^2 + 2\eta\langle w^{k+1} - w^k, G^k \rangle \leq 0 \\
 & \implies \frac{1}{2\eta}\|w^{k+1} - w^k\|_2^2 + \langle w^{k+1} - w^k, G^k \rangle \leq 0, \tag{22}
 \end{aligned}$$

where the third inequality holds since  $w^k$  is feasible for  $k \in \{1, \dots, K+1\}$ . Continuing from (22),

$$\begin{aligned} & f_\alpha(w^{k+1}) + \frac{1}{2\eta} \|w^{k+1} - w^k\|_2^2 + \langle w^{k+1} - w^k, G^k - \nabla f_\alpha(w^k) \rangle \\ & \leq f_\alpha(w^k) + \alpha^{-1} \sqrt{d} L_0 \|w^{k+1} - w^k\|_2^2 \end{aligned} \quad (23)$$

$$\begin{aligned} & \implies 2\eta^{-1} \langle w^{k+1} - w^k, G^k - \nabla f_\alpha(w^k) \rangle \\ & \leq 2\eta^{-1} (f_\alpha(w^k) - f_\alpha(w^{k+1})) - \eta^{-2} \|w^{k+1} - w^k\|_2^2 + 2\alpha^{-1} \sqrt{d} L_0 \eta^{-1} \|w^{k+1} - w^k\|_2^2, \end{aligned} \quad (24)$$

where the first inequality uses the smoothness of  $f_\alpha(w)$  (Nesterov, 2004, Lemma 1.2.3):

$$f_\alpha(w^{k+1}) \leq f_\alpha(w^k) + \langle \nabla f_\alpha(w^k), w^{k+1} - w^k \rangle + \alpha^{-1} \sqrt{d} L_0 \|w^{k+1} - w^k\|_2^2.$$

Plugging (24) into (21),

$$\begin{aligned} & \text{dist}(0, \nabla f_\alpha(w^{k+1}) + N(w^{k+1}, S \cap \bar{B}_\beta))^2 \\ & \leq (6\alpha^{-2} d L_0^2 + 4\alpha^{-1} \sqrt{d} L_0 \eta^{-1} + \eta^{-2}) \|w^{k+1} - w^k\|_2^2 + 3 \|\nabla f_\alpha(w^k) - G^k\|_2^2 \\ & \quad + 2\eta^{-1} (f_\alpha(w^k) - f_\alpha(w^{k+1})) - \eta^{-2} \|w^{k+1} - w^k\|_2^2 + 2\alpha^{-1} \sqrt{d} L_0 \eta^{-1} \|w^{k+1} - w^k\|_2^2 \\ & = (6\alpha^{-2} d L_0^2 + 6\alpha^{-1} \sqrt{d} L_0 \eta^{-1}) \|w^{k+1} - w^k\|_2^2 + 3 \|\nabla f_\alpha(w^k) - G^k\|_2^2 + 2\eta^{-1} (f_\alpha(w^k) - f_\alpha(w^{k+1})) \\ & = 6\alpha^{-1} \sqrt{d} L_0 (\alpha^{-1} \sqrt{d} L_0 + \eta^{-1}) \|w^{k+1} - w^k\|_2^2 + 3 \|\nabla f_\alpha(w^k) - G^k\|_2^2 + 2\eta^{-1} (f_\alpha(w^k) - f_\alpha(w^{k+1})). \end{aligned} \quad (25)$$

Rearranging (23) and using Young's inequality again for the second inequality,

$$\begin{aligned} & \left( \frac{1}{2\eta} - \alpha^{-1} \sqrt{d} L_0 \right) \|w^{k+1} - w^k\|_2^2 \\ & \leq f_\alpha(w^k) - f_\alpha(w^{k+1}) + \langle w^{k+1} - w^k, \nabla f_\alpha(w^k) - G^k \rangle \\ & \leq f_\alpha(w^k) - f_\alpha(w^{k+1}) + \frac{\alpha^{-1} \sqrt{d} L_0}{2} \|w^{k+1} - w^k\|_2^2 + \frac{\alpha}{2\sqrt{d} L_0} \|\nabla f_\alpha(w^k) - G^k\|_2^2 \\ & \implies \left( \frac{1}{2\eta} - \frac{3\alpha^{-1} \sqrt{d} L_0}{2} \right) \|w^{k+1} - w^k\|_2^2 \\ & \leq f_\alpha(w^k) - f_\alpha(w^{k+1}) + \frac{\alpha}{2\sqrt{d} L_0} \|\nabla f_\alpha(w^k) - G^k\|_2^2 \\ & \implies \frac{1}{2} (\theta + \eta^{-1} - 3\alpha^{-1} \sqrt{d} L_0) \|w^{k+1} - w^k\|_2^2 \\ & \leq f_\alpha(w^k) - f_\alpha(w^{k+1}) + \frac{\alpha}{2\sqrt{d} L_0} \|\nabla f_\alpha(w^k) - G^k\|_2^2 + \frac{\theta}{2} \|w^{k+1} - w^k\|_2^2 \end{aligned} \quad (26)$$

for an arbitrary  $\theta \in \mathbb{R}$ . Focusing on  $\|w^{k+1} - w^k\|_2^2$ ,

$$\begin{aligned} & w^{k+1} \in \operatorname{argmin}_{x \in \mathbb{R}^d} \{ \delta_{S \cap \bar{B}_\beta}(x) + \|x - (w^k - \eta G^k)\|_2 \} \\ & \implies \|w^{k+1} - (w^k - \eta G^k)\|_2 \leq \|w^k - (w^k - \eta G^k)\|_2 = \|\eta G^k\|_2. \end{aligned} \quad (27)$$

Using the reverse triangle inequality and (27),

$$\begin{aligned} \|w^{k+1} - w^k\| - \|\eta G^k\|_2 &\leq \|w^{k+1} - w^k + \eta G^k\|_2 \\ \implies \|w^{k+1} - w^k\| &\leq 2\|\eta G^k\|_2, \end{aligned}$$

and applying this bound in (26) assuming that  $\theta \geq 0$ ,

$$\begin{aligned} &\frac{1}{2}(\theta + \eta^{-1} - 3\alpha^{-1}\sqrt{d}L_0)\|w^{k+1} - w^k\|_2^2 \\ &\leq f_\alpha(w^k) - f_\alpha(w^{k+1}) + \frac{\alpha}{2\sqrt{d}L_0}\|\nabla f_\alpha(w^k) - G^k\|_2^2 + 2\theta\eta^2\|G^k\|_2^2. \end{aligned} \quad (28)$$

Assuming that  $\theta + \eta^{-1} - 3\alpha^{-1}\sqrt{d}L_0 > 0$  and plugging (28) into (25),

$$\begin{aligned} &\text{dist}(0, \nabla f_\alpha(w^{k+1}) + N(w^{k+1}, S \cap \bar{B}_\beta))^2 \\ &\leq \frac{6\alpha^{-1}\sqrt{d}L_0(\alpha^{-1}\sqrt{d}L_0 + \eta^{-1})}{\frac{1}{2}(\theta + \eta^{-1} - 3\alpha^{-1}\sqrt{d}L_0)}(f_\alpha(w^k) - f_\alpha(w^{k+1}) + \frac{\alpha}{2\sqrt{d}L_0}\|\nabla f_\alpha(w^k) - G^k\|_2^2 \\ &\quad + 2\theta\eta^2\|G^k\|_2^2) + 3\|\nabla f_\alpha(w^k) - G^k\|_2^2 + 2\eta^{-1}(f_\alpha(w^k) - f_\alpha(w^{k+1})) \\ \implies &\mathbb{E}[\text{dist}(0, \nabla f_\alpha(w^{k+1}) + N(w^{k+1}, S \cap \bar{B}_\beta))^2] \\ &\leq \frac{6\alpha^{-1}\sqrt{d}L_0(\alpha^{-1}\sqrt{d}L_0 + \eta^{-1})}{\frac{1}{2}(\theta + \eta^{-1} - 3\alpha^{-1}\sqrt{d}L_0)}(\mathbb{E}[f_\alpha(w^k) - f_\alpha(w^{k+1})] + \frac{\alpha}{2\sqrt{d}L_0}\mathbb{E}[\|\nabla f_\alpha(w^k) - G^k\|_2^2] \\ &\quad + 2\theta\eta^2\mathbb{E}[\|G^k\|_2^2]) + 3\mathbb{E}[\|\nabla f_\alpha(w^k) - G^k\|_2^2] + 2\eta^{-1}\mathbb{E}[f_\alpha(w^k) - f_\alpha(w^{k+1})] \\ \implies &\mathbb{E}[\text{dist}(0, \nabla f_\alpha(w^R) + N(w^R, S \cap \bar{B}_\beta))^2] \\ &\leq \frac{1}{K} \sum_{k=1}^K \left( \frac{6\alpha^{-1}\sqrt{d}L_0(\alpha^{-1}\sqrt{d}L_0 + \eta^{-1})}{\frac{1}{2}(\theta + \eta^{-1} - 3\alpha^{-1}\sqrt{d}L_0)}(\mathbb{E}[f_\alpha(w^k) - f_\alpha(w^{k+1})] + \frac{\alpha}{2\sqrt{d}L_0}\mathbb{E}[\|\nabla f_\alpha(w^k) - G^k\|_2^2] \right. \\ &\quad \left. + 2\theta\eta^2\mathbb{E}[\|G^k\|_2^2]) + 3\mathbb{E}[\|\nabla f_\alpha(w^k) - G^k\|_2^2] + 2\eta^{-1}\mathbb{E}[f_\alpha(w^k) - f_\alpha(w^{k+1})] \right) \\ &\leq \frac{6\alpha^{-1}\sqrt{d}L_0(\alpha^{-1}\sqrt{d}L_0 + \eta^{-1})}{\frac{1}{2}(\theta + \eta^{-1} - 3\alpha^{-1}\sqrt{d}L_0)} \left( \frac{\mathbb{E}[f_\alpha(w^1) - f_\alpha(w^{K+1})]}{K} + \frac{\alpha}{2\sqrt{d}L_0} \frac{vQ}{M} + 2\theta\eta^2 vQ \right) \\ &\quad + 3 \frac{vQ}{M} + \frac{2\eta^{-1}}{K} \mathbb{E}[f_\alpha(w^1) - f_\alpha(w^{K+1})] \\ &\leq \frac{6\alpha^{-1}\sqrt{d}L_0(\alpha^{-1}\sqrt{d}L_0 + \eta^{-1})}{\frac{1}{2}(\theta + \eta^{-1} - 3\alpha^{-1}\sqrt{d}L_0)} \left( \frac{f_\alpha(w^1) - f_\alpha(w^*)}{K} + \left( \frac{\alpha}{2\sqrt{d}L_0 M} + 2\theta\eta^2 \right) vQ \right) \\ &\quad + 3 \frac{vQ}{M} + \frac{2\eta^{-1}}{K} (f_\alpha(w^1) - f_\alpha(w^*)) \\ &\leq \frac{6\alpha^{-1}\sqrt{d}L_0(\alpha^{-1}\sqrt{d}L_0 + \eta^{-1})}{\frac{1}{2}(\theta + \eta^{-1} - 3\alpha^{-1}\sqrt{d}L_0)} \left( \frac{\Delta}{K} + \left( \frac{\alpha}{2\sqrt{d}L_0 M} + 2\theta\eta^2 \right) vQ \right) \\ &\quad + 3 \frac{vQ}{M} + \frac{2\eta^{-1}\Delta}{K}, \end{aligned} \quad (29)$$

where the third last inequality uses Proposition 7 and the last inequality uses that  $\Delta \geq f_\alpha(w^1) - f_\alpha(w^*)$ . It holds that  $f_\alpha(w^1) - f_\alpha(w^*) \leq 2\beta\sqrt{d}L_0$  from Proposition 5.4 and

$f_\alpha(w^1) - f_\alpha(w_\alpha^*) \leq f(w^1) - f_\alpha(w_\alpha^*) + \alpha L_0 \sqrt{\frac{d}{12}} \leq f(w^1) - f(w^*) + \alpha L_0 \sqrt{\frac{d}{3}}$  using Propositions 5.2 and 5.3.

Let  $\theta = 3\tau\alpha^{-1}\sqrt{d}L_0$  and  $\eta^{-1} = 3\rho\alpha^{-1}\sqrt{d}L_0$  for  $\tau \geq 0$  and  $\rho > 0$  such that  $\tau + \rho > 1$ . Continuing from (29),

$$\begin{aligned}
 & \mathbb{E}[\text{dist}(0, \nabla f_\alpha(w^R) + N(w^R, S \cap \bar{B}_\beta))^2] \\
 & \leq \frac{6\alpha^{-1}\sqrt{d}L_0(\alpha^{-1}\sqrt{d}L_0 + 3\rho\alpha^{-1}\sqrt{d}L_0)}{\frac{1}{2}(\tau + \rho - 1)3\alpha^{-1}\sqrt{d}L_0} \left( \frac{\Delta}{K} + \left( \frac{\alpha}{2\sqrt{d}L_0M} + \frac{\tau}{\rho^2} \frac{2\alpha}{3\sqrt{d}L_0} \right) vQ \right) \\
 & \quad + 3 \frac{vQ}{M} + \frac{6\rho\alpha^{-1}\sqrt{d}L_0\Delta}{K} \\
 & = \frac{6\alpha^{-2}dL_0^2(1+3\rho)}{\frac{1}{2}(\tau + \rho - 1)3\alpha^{-1}\sqrt{d}L_0} \left( \frac{\Delta}{K} + \left( \frac{1}{2M} + \frac{\tau}{\rho^2} \frac{2}{3} \right) \frac{\alpha}{\sqrt{d}L_0} vQ \right) + 3 \frac{vQ}{M} + \frac{6\rho\alpha^{-1}\sqrt{d}L_0\Delta}{K} \\
 & = \frac{4\alpha^{-1}\sqrt{d}L_0(1+3\rho)}{(\tau + \rho - 1)} \left( \frac{\Delta}{K} + \left( \frac{1}{2M} + \frac{\tau}{\rho^2} \frac{2}{3} \right) \frac{\alpha}{\sqrt{d}L_0} vQ \right) + 3 \frac{vQ}{M} + \frac{6\rho\alpha^{-1}\sqrt{d}L_0\Delta}{K} \\
 & = \left( \frac{2(1+3\rho)}{(\tau + \rho - 1)} + 3\rho \right) \frac{2\alpha^{-1}\sqrt{d}L_0\Delta}{K} + \left( \frac{4(1+3\rho)}{(\tau + \rho - 1)} \left( \frac{1}{2} + \frac{2}{3} \frac{M\tau}{\rho^2} \right) + 3 \right) \frac{vQ}{M}.
 \end{aligned}$$

■

**Proof of Corollary 17** In order for  $|f_\alpha(w) - f(w)| \leq \epsilon_1$  for all  $w \in \bar{B}_\beta$ , we require  $\alpha \leq \frac{\epsilon_1}{L_0\sqrt{\frac{d}{12}}}$  from Proposition 5.2, and for  $\hat{\alpha} \leq \epsilon_3$ , we require  $\alpha \leq \frac{2\epsilon_3}{\sqrt{d}}$  from Corollary 16. To ensure that  $\mathbb{E}[\text{dist}(0, \nabla f_\alpha(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta))] \leq \epsilon$  or  $\mathbb{E}[\text{dist}(0, \bar{\partial} f_{\hat{\alpha}}(w^R) + N(w^R, S \cap \bar{B}_\beta))] \leq \epsilon$  for  $\epsilon = \epsilon_2$  or  $\epsilon_4$ , using (9) or (11) and Jensen's inequality, it is sufficient for

$$C_1 \frac{2\alpha^{-1}\sqrt{d}L_0\Delta}{K} + C_2 \frac{vQ}{M} \leq \epsilon^2.$$

Taking  $y \in (0, 1)$ , we choose  $K$  and  $M$  such that

$$C_1 \frac{2\alpha^{-1}\sqrt{d}L_0\Delta}{K} \leq y\epsilon^2$$

and

$$C_2 \frac{vQ}{M} \leq (1-y)\epsilon^2,$$

which results in requiring

$$C_1 \frac{2\alpha^{-1}\sqrt{d}L_0\Delta}{y\epsilon^2} \leq K \tag{30}$$

and

$$C_2 \frac{vQ}{(1-y)\epsilon^2} \leq M.$$

Assuming SPA is run for the full  $K$  iterations, the number of gradient calls will equal  $KM$ . Considering the bound on KM,

$$C_1 \frac{2\alpha^{-1}\sqrt{d}L_0\Delta}{y\epsilon^2} C_2 \frac{vQ}{(1-y)\epsilon^2} \leq KM, \quad (31)$$

minimizing the left-hand side of (31) in terms of  $y$  gives  $y = 0.5$ , and minimizing the left-hand side of (30) in terms of  $\alpha$  sets  $\alpha = \frac{\epsilon_1}{L_0\sqrt{\frac{d}{12}}}$  or  $\alpha = \frac{2\epsilon_3}{\sqrt{d}}$ . Using these values for  $\alpha$  and  $y$  gives the bounds for  $K$  of

$$C_1 \sqrt{\frac{4}{3}} \frac{dL_0^2\Delta}{\epsilon_1\epsilon_2^2} \leq K$$

for an  $(\epsilon_1, \epsilon_2)$ -solution, and

$$C_1 \frac{2dL_0\Delta}{\epsilon_3\epsilon_4^2} \leq K$$

for an  $(\epsilon_3, \epsilon_4)$ -solution. The bound for  $M$  equals

$$C_2 \frac{2vQ}{\epsilon^2} \leq M$$

for  $\epsilon = \epsilon_2$  or  $\epsilon_4$ .

Taking the choices of  $K$  and  $M$  given in this corollary and the upper bound of  $KM$  for the total number of gradient calls gives the gradient call complexities of  $O(\epsilon_1^{-1}\epsilon_2^{-4})$  and  $O(\epsilon_3^{-1}\epsilon_4^{-4})$  to achieve an expected  $(\epsilon_1, \epsilon_2)$  and  $(\epsilon_3, \epsilon_4)$ -stationary point, respectively. Given that one projection is done per iteration, the projection operator complexities are  $O(\epsilon_1^{-1}\epsilon_2^{-2})$  and  $O(\epsilon_3^{-1}\epsilon_4^{-2})$ .  $\blacksquare$

**Proof of Corollary 18** Following (Ghadimi and Lan, 2013, Equation 2.28),

$$\begin{aligned} & \text{dist}(0, G(w^*) + N(w^*, S \cap \bar{B}_\beta))^2 \\ &= \min_{w \in W} \text{dist}(0, G(w) + N(w, S \cap \bar{B}_\beta))^2 \\ &= \min_{w \in W} \text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \bar{B}_\beta) + G(w) - \nabla f_\alpha(w))^2 \\ &\leq \min_{w \in W} \{2 \text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \bar{B}_\beta))^2 + 2\|G(w) - \nabla f_\alpha(w)\|_2^2\} \\ &\leq \min_{w \in W} 2 \text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \bar{B}_\beta))^2 + \max_{w \in W} 2\|G(w) - \nabla f_\alpha(w)\|_2^2, \end{aligned} \quad (32)$$

where the first inequality holds since

$$\begin{aligned} & \text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \bar{B}_\beta) + G(w) - \nabla f_\alpha(w))^2 \\ &= \min_{\nu \in N(w, S \cap \bar{B}_\beta)} \|\nabla f_\alpha(w) + \nu + G(w) - \nabla f_\alpha(w)\|_2^2 \\ &= \min_{\nu \in N(w, S \cap \bar{B}_\beta)} \|\nabla f_\alpha(w) + \nu\|_2^2 + 2\langle \nabla f_\alpha(w) + \nu, G(w) - \nabla f_\alpha(w) \rangle + \|G(w) - \nabla f_\alpha(w)\|_2^2 \\ &\leq \min_{\nu \in N(w, S \cap \bar{B}_\beta)} 2\|\nabla f_\alpha(w) + \nu\|_2^2 + 2\|G(w) - \nabla f_\alpha(w)\|_2^2 \end{aligned}$$



using Young's inequality. Using Young's inequality again for the first inequality and (32) for the second inequality,

$$\begin{aligned}
 & \text{dist}(0, \nabla f_\alpha(w^*) + N(w^*, S \cap \bar{B}_\beta))^2 \\
 \leq & 2 \text{dist}(0, G(w^*) + N(w^*, S \cap \bar{B}_\beta))^2 + 2 \|\nabla f_\alpha(w^*) - G(w^*)\|_2^2 \\
 \leq & 4 \min_{w \in W} \text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \bar{B}_\beta))^2 + 4 \max_{w \in W} \|G(w) - \nabla f_\alpha(w)\|_2^2 + 2 \|\nabla f_\alpha(w^*) - G(w^*)\|_2^2 \\
 \leq & 4 \min_{w \in W} \text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \bar{B}_\beta))^2 + 6 \max_{w \in W} \|G(w) - \nabla f_\alpha(w)\|_2^2. \tag{33}
 \end{aligned}$$

Let the right-hand side of (9) be denoted as

$$D := C_1 \frac{2\alpha^{-1} \sqrt{d} L_0 \Delta}{K} + C_2 \frac{vQ}{M}.$$

Considering the first term of (33),

$$\begin{aligned}
 & \mathbb{P} \left( 4 \min_{w \in W} \text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \bar{B}_\beta))^2 \geq 4eD \right) \\
 = & \mathbb{P} \left( \bigcap_{i=1}^r \mathbb{P} (4 \text{dist}(0, \nabla f_\alpha(w^i) + N(w^i, S \cap \bar{B}_\beta))^2 \geq 4eD) \right) \\
 \leq & e^{-r}
 \end{aligned}$$

using Markov's inequality. For the second term of (33), using Proposition 7 and Boole's inequality,

$$\begin{aligned}
 & \mathbb{P} \left( 6 \max_{w \in W} \|G(w) - \nabla f_\alpha(w)\|_2^2 \geq 6\psi \frac{vQ}{T} \right) \\
 = & \mathbb{P} \left( \bigcup_{i=1}^r \left\{ 6 \|G(w^i) - \nabla f_\alpha(w^i)\|_2^2 \geq 6\psi \frac{vQ}{T} \right\} \right) \\
 \leq & \frac{r}{\psi}.
 \end{aligned}$$

Combining these two probability inequalities together, for the left-hand side of (33),

$$\begin{aligned}
 & \mathbb{P} \left( \text{dist}(0, \nabla f_\alpha(w^*) + N(w^*, S \cap \bar{B}_\beta))^2 \geq 4eD + 6\psi \frac{vQ}{T} \right) \\
 \leq & \mathbb{P} \left( 4 \min_{w \in W} \text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \bar{B}_\beta))^2 + 6 \max_{w \in W} \|G(w) - \nabla f_\alpha(w)\|_2^2 \geq 4eD + 6\psi \frac{vQ}{T} \right) \\
 \leq & \mathbb{P} \left( \left\{ 4 \min_{w \in W} \text{dist}(0, \nabla f_\alpha(w) + N(w, S \cap \bar{B}_\beta))^2 \geq 4eD \right\} \right. \\
 & \left. \cup \left\{ 6 \max_{w \in W} \|G(w) - \nabla f_\alpha(w)\|_2^2 \geq 6\psi \frac{vQ}{T} \right\} \right) \\
 \leq & e^{-r} + \frac{r}{\psi}. \tag{34}
 \end{aligned}$$

An upper bound on the total number of gradient calls required for computing  $W$  and  $G(w)$  for  $w \in W$  is equal to  $r(KM + T)$ . Using (34), the minimization of  $r(KM + T)$

while ensuring that  $\mathbb{P}(\text{dist}(0, \nabla f_\alpha(w^*) + N(w^*, S \cap \bar{B}_\beta)) > \epsilon_{2/4}) \leq \gamma$ , where  $\alpha = \frac{\epsilon_1}{L_0 \sqrt{\frac{d}{12}}}$  and  $\epsilon_{2/4} = \epsilon_2$ , or  $\alpha = \frac{2\epsilon_3}{\sqrt{d}}$  and  $\epsilon_{2/4} = \epsilon_4$  (assuming that  $\kappa > \beta + \epsilon_3$ ) can be written as

$$\begin{aligned} & \min_{r, K, M, T, \psi} r(KM + T) \\ & \text{s.t. } 4e \left( C_1 \frac{2\alpha^{-1} \sqrt{d} L_0 \Delta}{K} + C_2 \frac{vQ}{M} \right) + 6\psi \frac{vQ}{T} \leq \epsilon_{2/4}^2 \\ & e^{-r} + \frac{r}{\psi} \leq \gamma \\ & r, K, M, T \in \mathbb{Z}_{>0}, \quad \psi > 0. \end{aligned} \quad (35)$$

Writing (35) as

$$C_1 \frac{2\alpha^{-1} \sqrt{d} L_0 \Delta}{K} + C_2 \frac{vQ}{M} \leq \frac{\epsilon_{2/4}^2 - 6\psi \frac{vQ}{T}}{4e},$$

we set  $K$  and  $M$  according to Corollary 17 to find an expected  $(\epsilon_1, \epsilon'_2)$  or  $(\epsilon_3, \epsilon'_4)$ -stationary point for  $\epsilon'_2 = \sqrt{\frac{\epsilon_2^2 - 6\psi \frac{vQ}{T}}{4e}}$  or  $\epsilon'_4 = \sqrt{\frac{\epsilon_4^2 - 6\psi \frac{vQ}{T}}{4e}}$  assuming that  $\epsilon_{2/4}^2 - 6\psi \frac{vQ}{T} > 0$ :

$$K^* = \left\lceil C_1 \frac{2}{\chi} \frac{dL_0^2 \Delta}{\epsilon_{1/3} (\epsilon'_{2/4})^2} \right\rceil \quad \text{and} \quad M^* = \left\lceil C_2 \frac{2vQ}{(\epsilon'_{2/4})^2} \right\rceil,$$

where  $\epsilon_{1/3} = \epsilon_1$  or  $\epsilon_3$ ,  $\epsilon'_{2/4} = \epsilon'_2$  or  $\epsilon'_4$ , and  $\chi = \sqrt{3}$  or  $L_0$  for an expected  $(\epsilon_1, \epsilon'_2)$  or  $(\epsilon_3, \epsilon'_4)$ -stationary point, respectively. The optimization problem then becomes

$$\begin{aligned} & \min_{r, T, \psi} r(K^*M^* + T) \\ & \text{s.t. } e^{-r} + \frac{r}{\psi} \leq \gamma \end{aligned} \quad (36)$$

$$\begin{aligned} & \epsilon_{2/4}^2 - 6\psi \frac{vQ}{T} > 0 \\ & r, T \in \mathbb{Z}_{>0}, \quad \psi > 0. \end{aligned} \quad (37)$$

For any  $c \in (0, 1)$ , let

$$e^{-r} \leq c\gamma \quad \text{and} \quad \frac{r}{\psi} \leq (1-c)\gamma.$$

Setting  $r$  and  $\psi$  to  $r^* = \lceil -\ln(c\gamma) \rceil$  and  $\psi^* = \frac{\lceil -\ln(c\gamma) \rceil}{(1-c)\gamma}$  is then valid for (36). For any  $\phi > 1$ , setting  $T$  to  $T^* = \lceil 6\phi\psi^* \frac{vQ}{\epsilon_{2/4}^2} \rceil$  is valid for (37). The total number of gradient calls then

equals

$$\begin{aligned}
 & r^* \left( K^* M^* + T^* \right) \\
 &= r^* \left( \left[ C_1 \frac{2}{\chi} \frac{dL_0^2 \Delta}{\epsilon_{1/3} (\epsilon'_{2/4})^2} \right] \left[ C_2 \frac{2vQ}{(\epsilon'_{2/4})^2} \right] + T^* \right) \\
 &= r^* \left( \left[ C_1 \frac{2}{\chi} \frac{dL_0^2 \Delta}{\epsilon_{1/3} \frac{\epsilon_{2/4}^2 - 6\psi^* \frac{vQ}{T^*}}{4e}} \right] \left[ C_2 \frac{2vQ}{\frac{\epsilon_{2/4}^2 - 6\psi^* \frac{vQ}{T^*}}{4e}} \right] + T^* \right) \\
 &\leq r^* \left( \left[ C_1 \frac{2}{\chi} \frac{4edL_0^2 \Delta}{\epsilon_{1/3} \epsilon_{2/4}^2 (1 - \phi^{-1})} \right] \left[ C_2 \frac{8evQ}{\epsilon_{2/4}^2 (1 - \phi^{-1})} \right] + T^* \right) \\
 &= [-\ln(c\gamma)] \left( \left[ C_1 \frac{2}{\chi} \frac{4edL_0^2 \Delta}{\epsilon_{1/3} \epsilon_{2/4}^2 (1 - \phi^{-1})} \right] \left[ C_2 \frac{8evQ}{\epsilon_{2/4}^2 (1 - \phi^{-1})} \right] + \left[ 6\phi \frac{[-\ln(c\gamma)] vQ}{(1-c)\gamma \epsilon_{2/4}^2} \right] \right),
 \end{aligned}$$

where the inequality holds since

$$\epsilon_{2/4}^2 - 6\psi^* \frac{vQ}{T^*} = \epsilon_{2/4}^2 - 6\psi^* \frac{vQ}{\left\lceil 6\phi\psi^* \frac{vQ}{\epsilon_{2/4}^2} \right\rceil} \geq \epsilon_{2/4}^2 - 6\psi^* \frac{vQ}{6\phi\psi^* \frac{vQ}{\epsilon_{2/4}^2}} = \epsilon_{2/4}^2 (1 - \phi^{-1}),$$

and the gradient call complexity equals  $\tilde{O}(\epsilon_1^{-1} \epsilon_2^{-4} + \gamma^{-1} \epsilon_2^{-2})$  or  $\tilde{O}(\epsilon_3^{-1} \epsilon_4^{-4} + \gamma^{-1} \epsilon_4^{-2})$  for an  $(\epsilon_1, \epsilon_2)$  or  $(\epsilon_3, \epsilon_4)$ -stationary point with probability  $1 - \gamma$ . The number of projection computations is upper bounded by  $r^* K^*$  which has a complexity of  $\tilde{O}(\epsilon_1^{-1} \epsilon_2^{-2})$  or  $\tilde{O}(\epsilon_3^{-1} \epsilon_4^{-2})$ . ■

**Proof of Proposition 19** For simplicity let  $I := I(w)$ . The distance function

$$\text{dist}(0, G(w) + N(w, C \cap \bar{B}_\beta)) = \min_{v \in N(w, C \cap \bar{B}_\beta)} \|G(w) + v\|_2.$$

When  $Y = I$ , we only need to focus on two subsets of  $[n]$ :  $I$  and  $[n] \setminus I$ . Each subset of elements  $v^i$  for  $i \in I$  is set to its optimal value while following Proposition 12: when  $w_j^i = \beta$ , it is required that  $v_j^i \geq 0$ , hence it is optimal to set  $v_j^i = 0$  when  $G_j^i(w) \geq 0$  and to set  $v_j^i = -G_j^i(w)$  when  $G_j^i(w) < 0$ . Similarly, when  $w_j^i = -\beta$ , it is required that  $v_j^i \leq 0$ , and it is optimal to set  $v_j^i = 0$  when  $G_j^i(w) \leq 0$  and to set  $v_j^i = -G_j^i(w)$  when  $G_j^i(w) > 0$ . When  $i \in I$  and  $|w_j^i| < \beta$ , it is required that  $v_j^i = 0$ . When  $i \notin I$ ,  $v_j^i$  is a free variable which is optimally set to  $v_j^i = -G_j^i(w)$  for all  $j \in [d_i]$ .

When there exists an  $X \in Y$  such that  $X \setminus I(w) \neq \{\emptyset\}$ , and assuming that  $\{p_i\} \subset \mathbb{Q}_{>0}$ ,  $C$  can be written equivalently with parameters  $\{p'_i\} = \{cp_i\}$  and  $m' = cm$  for a  $c \in \mathbb{Z}_{>0}$  sufficiently large such that  $\{p'_i\} \subset \mathbb{Z}_{>0}$ , so without loss of generality we can assume that  $\{p_i\} \subset \mathbb{Z}_{>0}$ .

The distance function has the same optimal solutions as

$$\min_{v \in N(w, C \cap \bar{B}_\beta)} \|G(w) + v\|_2^2,$$

which can be written as

$$\begin{aligned}
 & \min \|G(w) + v\|_2^2 \\
 & \text{s.t. } \sum_{i \in I} p_i + \sum_{i \notin I} z_i p_i \leq m \\
 & \quad \sum_{i \in I} p_i + \sum_{i \notin I} z_i p_i + (1 - z_j) p_j \geq (1 - z_j)(\lfloor m \rfloor + 1) \quad \forall j \notin I \\
 & \quad v_j^i = \begin{cases} -G_j^i(w) & \text{if } U_j^i \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in I, \forall j \in [d_i] \\
 & \quad v_j^i = -G_j^i(w)(1 - z_i) \quad \forall i \notin I, \forall j \in [d_i] \\
 & \quad z_i \in \{0, 1\} \quad \forall i \notin I.
 \end{aligned}$$

Using Proposition 12, the first two constraints determine an  $X$  equal to  $I$  and the indices for which  $z_i = 1$ , where the second constraint enforces that  $\sum_{i \in I} p_i + \sum_{i \notin I} z_i p_i + p_j > m$ , given the integrality of  $\{p_i\}$ , for  $j \notin X$ , i.e.  $z_j = 0$ . The third and fourth constraints set each subset of elements  $v^i$  for  $i \in I$  and  $i \notin I$  to their optimal value, respectively. For the fourth constraint, when  $z_i = 1$ , i.e.  $i \in X$ ,  $v^i$  must be set to  $v^i = 0$  given that  $w^i = 0$ , and when  $z_i = 0$ ,  $v^i$  is free to be chosen as  $v^i = -G^i(w)$  to minimize the objective.

The final binary integer program to compute  $\text{dist}(0, G(w) + N(w, C \cap \bar{B}_\beta))$  is as follows, removing the  $v$  decision variables.

$$\begin{aligned}
 & \min \sum_{i \notin I} \|G^i(w)\|_2^2 z_i \tag{38} \\
 & \text{s.t. } \sum_{i \in I} p_i + \sum_{i \notin I} z_i p_i \leq m \\
 & \quad \sum_{i \in I} p_i + \sum_{i \notin I} z_i p_i + (1 - z_j) p_j \geq (1 - z_j)(\lfloor m \rfloor + 1) \quad \forall j \notin I \\
 & \quad z_i \in \{0, 1\} \quad \forall i \notin I.
 \end{aligned}$$

Given an optimal solution  $z^*$  to (38), let  $y^*$  be defined as

$$\begin{aligned}
 (y_j^i)^* &= \begin{cases} 0 & \text{if } U_j^i \\ 1 & \text{otherwise} \end{cases} \quad \forall i \in I, \forall j \in [d_i] \\
 (y_j^i)^* &= z_i^* \quad \forall i \notin I, \forall j \in [d_i].
 \end{aligned}$$

It follows that  $\text{dist}(0, G(w) + N(w, C \cap \bar{B}_\beta)) = \sqrt{\sum_{i \in [n]} \sum_{j \in d_i} (G_j^i(w))^2 (y_j^i)^*}$ . ■

**Proof of Theorem 20** If  $\beta$  is finite,  $\{w^i\} \subset \bar{B}_\beta$  is a bounded sequence and there exists an accumulation point  $\bar{w}$  of  $\{w^i\}$ . Otherwise, assume there exists an accumulation point  $\bar{w}$  of  $\{w^i\}$ . For simplicity, let  $\{w^i\}$  be redefined as a subsequence of  $\{w^i\}$  such that  $\lim_{i \rightarrow \infty} w^i = \bar{w}$ .

Since  $\|\zeta\|_2 \leq L_0$  for all  $\zeta \in \bar{\partial}f(w)$  for all  $w \in \bar{B}_\beta$  (Clarke, 1990, Proposition 2.1.2) and  $0 \in N(w, S \cap \bar{B}_\beta)$ ,

$$\text{dist}(0, \bar{\partial}f(w) + N(w, S \cap \bar{B}_\beta)) = \text{dist}(0, \bar{\partial}f(w) + N(w, S \cap \bar{B}_\beta) \cap \bar{B}(0, 2L_0))$$

for  $w \in \bar{B}_\beta$ : If  $\zeta \in \bar{\partial}f(w)$ ,  $\nu \in N(w, S \cap \bar{B}_\beta)$ , and  $\|\nu\|_2 > 2L_0$ , then by the reverse triangle inequality,  $\|\zeta + 0\|_2 \leq L_0 < \|\nu\|_2 - \|\zeta\|_2 \leq \|\nu + \zeta\|_2$ .

Given that  $N(w, S \cap \bar{B}_\beta)$  is outer semicontinuous (proof of Proposition 14), for all  $\omega_1 > 0$ , there exists an  $\omega_2 > 0$  such that

$$N(\hat{w}, S \cap \bar{B}_\beta) \cap \bar{B}(0, 2L_0) \subseteq N(w, S \cap \bar{B}_\beta) + \bar{B}(0, \omega_1) \quad (39)$$

for all  $\hat{w} \in \bar{B}(w, \omega_2)$  (Rockafellar and Wets, 2009, Proposition 5.12).

For any  $i \in \mathbb{N}$  there exists an  $I \in \mathbb{N}$  such that for all  $j > I$ ,  $w^j$  is an expected  $(\frac{\epsilon_3^i}{2}, \frac{\epsilon_4^i}{2i})$ -stationary point and  $\|w^j - \bar{w}\|_2 \leq \min(\frac{\epsilon_3^i}{2}, \omega_2^i)$ , where  $\omega_2^i > 0$  is chosen such that (39) holds for  $\omega_1 = \frac{\epsilon_4^i}{2i}$ ,  $\omega_2 = \omega_2^i$ , and  $w = \bar{w}$ . For such  $w^j$ ,  $\bar{B}(w^j, \frac{\epsilon_3^i}{2}) \subset \bar{B}(\bar{w}, \epsilon_3^i)$ , hence

$$\{\bar{\partial}f(w) : w \in \bar{B}(w^j, \epsilon_3^i/2)\} \subseteq \{\bar{\partial}f(w) : w \in \bar{B}(\bar{w}, \epsilon_3^i)\},$$

and  $\bar{\partial}_{\epsilon_3^i/2}f(w^j) \subseteq \bar{\partial}_{\epsilon_3^i}f(\bar{w})$ . In addition,

$$\begin{aligned} \frac{\epsilon_4^i}{2i} &\geq \mathbb{E}[\text{dist}(0, \bar{\partial}_{\epsilon_3^i/2}f(w^j) + N(w^j, S \cap \bar{B}_\beta))] \\ &= \mathbb{E}[\text{dist}(0, \bar{\partial}_{\epsilon_3^i/2}f(w^j) + N(w^j, S \cap \bar{B}_\beta) \cap \bar{B}(0, 2L_0))] \\ &\geq \mathbb{E}[\text{dist}(0, \bar{\partial}_{\epsilon_3^i}f(\bar{w}) + N(w^j, S \cap \bar{B}_\beta) \cap \bar{B}(0, 2L_0))] \\ &\geq \mathbb{E}[\text{dist}(0, \bar{\partial}_{\epsilon_3^i}f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) + \bar{B}(0, \epsilon_4^i/(2i)))] \end{aligned} \quad (40)$$

Using the reverse triangle inequality for the first inequality,

$$\begin{aligned} &\text{dist}(0, \bar{\partial}_{\epsilon_3^i}f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) + \bar{B}(0, \epsilon_4^i/(2i))) \\ &= \min_{\substack{z \in \bar{\partial}_{\epsilon_3^i}f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) \\ y \in \bar{B}(0, \epsilon_4^i/(2i))}} \|z + y\|_2 \\ &\geq \min_{\substack{z \in \bar{\partial}_{\epsilon_3^i}f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) \\ y \in \bar{B}(0, \epsilon_4^i/(2i))}} \|z\|_2 - \|y\|_2 \\ &= \text{dist}(0, \bar{\partial}_{\epsilon_3^i}f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta)) - \frac{\epsilon_4^i}{2i} \\ &\implies \mathbb{E}[\text{dist}(0, \bar{\partial}_{\epsilon_3^i}f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) + \bar{B}(0, \epsilon_4^i/(2i)))] \\ &\geq \mathbb{E}[\text{dist}(0, \bar{\partial}_{\epsilon_3^i}f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta))] - \frac{\epsilon_4^i}{2i}. \end{aligned} \quad (41)$$

Applying (41) in (40),

$$\mathbb{E}[\text{dist}(0, \bar{\partial}_{\epsilon_3^i} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta))] \leq \frac{\epsilon_4^i}{i}.$$

From Markov's inequality,

$$\mathbb{P}[\text{dist}(0, \bar{\partial}_{\epsilon_3^i} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) \geq \frac{1}{i}] \leq \epsilon_4^i.$$

The sets

$$V_i := \{\bar{w} \in \mathbb{R}^d : \text{dist}(0, \bar{\partial}_{\epsilon_3^i} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) \geq \frac{1}{i}\}$$

are monotonically increasing:  $V_i \subseteq V_{i+1}$ , as  $\text{dist}(0, \bar{\partial}_{\epsilon_3^i} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) \leq \text{dist}(0, \bar{\partial}_{\epsilon_3^{i+1}} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta)$  and  $\frac{1}{i} > \frac{1}{i+1}$ . The limit  $\lim_{i \rightarrow \infty} V_i = \bigcup_{i \geq 1} V_i$  exists (Bartle, 1995, Excecise 2.F.), and is Borel measurable as a countable union of measurable sets, given that the functions  $\text{dist}(0, \bar{\partial}_{\epsilon_3^i} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta)$  are Borel measurable from Proposition 14.

We now want to prove that  $\lim_{i \rightarrow \infty} V_i = \{\bar{w} \in \mathbb{R}^d : \text{dist}(0, \bar{\partial} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) > 0\}$ . For any  $w \in \bigcup_{i \geq 1} V_i$  there exists an  $i \geq 1$  such that  $\text{dist}(0, \bar{\partial} f(w) + N(w, S \cap \bar{B}_\beta) \geq \text{dist}(0, \bar{\partial}_{\epsilon_3^i} f(w) + N(w, S \cap \bar{B}_\beta) \geq \frac{1}{i} > 0$ , hence  $w \in \{\bar{w} \in \mathbb{R}^d : \text{dist}(0, \bar{\partial} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) > 0\}$ .

For a  $w \in \{\bar{w} \in \mathbb{R}^d : \text{dist}(0, \bar{\partial} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) > 0\}$ , let  $\omega = \text{dist}(0, \bar{\partial} f(w) + N(w, S \cap \bar{B}_\beta)$ . As was shown with  $N(w, S \cap \bar{B}_\beta)$ , given that the Clarke subdifferential is an upper semicontinuous set valued mapping (Clarke, 1990, Proposition 2.1.5 (d)), for all  $\omega_1 > 0$ , there exists an  $\omega_2 > 0$  such that  $\bar{\partial} f(\hat{w}) \subseteq \bar{\partial} f(w) + \bar{B}(0, \omega_1)$  for all  $\hat{w} \in \bar{B}(w, \omega_2)$ , from which it follows that  $\bar{\partial}_{\omega_2} f(w) \subseteq \text{co}\{\bar{\partial} f(w) + \bar{B}(0, \omega_1)\} = \bar{\partial} f(w) + \bar{B}(0, \omega_1)$ , given that  $\bar{\partial} f(w)$  is convex, and

$$\text{dist}(0, \bar{\partial}_{\omega_2} f(w) + N(w, S \cap \bar{B}_\beta)) \geq \text{dist}(0, \bar{\partial} f(w) + \bar{B}(0, \omega_1) + N(w, S \cap \bar{B}_\beta)). \quad (42)$$

Just as in proving (41),  $\text{dist}(0, \bar{\partial} f(w) + \bar{B}(0, \omega_1) + N(w, S \cap \bar{B}_\beta))$  can be bounded below:

$$\text{dist}(0, \bar{\partial} f(w) + \bar{B}(0, \omega_1) + N(w, S \cap \bar{B}_\beta)) \geq \omega - \omega_1. \quad (43)$$

Choosing  $\omega_1 = \frac{\omega}{2}$ , there exists an  $\omega_2 > 0$  such that  $\text{dist}(0, \bar{\partial}_{\omega_2} f(w) + N(w, S \cap \bar{B}_\beta)) \geq \frac{\omega}{2}$  from (42) and (43). A  $J \in \mathbb{N}$  exists such that for all  $i \geq J$ ,  $\epsilon_3^i \leq \omega_2$ . Setting  $I \geq \max\{J, \frac{2}{\omega}\}$ ,

$$\text{dist}(0, \bar{\partial}_{\epsilon_3^i} f(w) + N(w, S \cap \bar{B}_\beta)) \geq \frac{1}{i},$$

i.e.  $w \in V_i$ , for all  $i \geq I$ , proving that  $\lim_{i \rightarrow \infty} V_i = \{\bar{w} \in \mathbb{R}^d : \text{dist}(0, \bar{\partial} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta) > 0\}$ .

It holds that  $\mathbb{P}[\text{dist}(0, \bar{\partial}f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta)) = 0] = 1$  as

$$\begin{aligned} \mathbb{P}[\text{dist}(0, \bar{\partial}f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta)) > 0] &= \lim_{i \rightarrow \infty} \mathbb{P}[\text{dist}(0, \bar{\partial}_{\epsilon_3^i} f(\bar{w}) + N(\bar{w}, S \cap \bar{B}_\beta)) \geq \frac{1}{i}] \\ &\leq \lim_{i \rightarrow \infty} \epsilon_4^i = 0, \end{aligned}$$

where the equality holds given that  $V_i \subseteq V_{i+1}$  (Shreve, 2004, Theorem A.1.1).  $\blacksquare$

## Appendix E. Section 8 Proofs

**Proof of Proposition 21** For all  $j \in I_k$ ,  $\bar{\partial}\sigma_j(\cdot)$  is outer semicontinuous (Clarke, 1990, Proposition 2.1.5 (d); Rockafellar and Wets, 2009, Theorem 5.19), hence measurable (Rockafellar and Wets, 2009, Exercise 14.9). There then exists a measurable selection  $\tilde{\nabla}\sigma_j(\cdot) \in \bar{\partial}\sigma_j(\cdot)$  for all  $j \in I_k$  (Rockafellar and Wets, 2009, Theorem 14.6).

As the product and sum of real-valued Borel measurable functions, see (Bolte and Pauwels, 2021, Algorithm 3),  $\tilde{\nabla}F(w, \xi_k)$  is Borel measurable in  $w \in \mathbb{R}^d$  for each  $\xi_k \in \{\xi_i\}_{i=1}^\infty$ . By the assumption that for each  $i \in \{1, 2, \dots\}$ , all  $\{\sigma_j\}_{j \in I_i}$  are definable in the same o-minimal structure, the Clarke subdifferentials  $\{\bar{\partial}\sigma_j\}_{j \in I_i}$  are definable conservative fields in said o-minimal structure as well (Bolte and Pauwels, 2021, Remark 8), hence for each  $\xi_k \in \{\xi_i\}_{i=1}^\infty$ ,  $\tilde{\nabla}F(w, \xi_k)$  will equal the gradient of  $F(w, \xi_k)$  for almost every  $w \in \mathbb{R}^d$  following (Bolte and Pauwels, 2021, Corollary 5). Further,  $\tilde{\nabla}F(w, \xi)$  will equal the gradient of  $F(w, \xi)$  for all  $(w, \xi) \in \mathbb{R}^{d+p}$  except on a countable number of null sets: The set  $\{(w, \xi) : w \in \mathbb{R}^d, \xi \notin \{\xi_i\}_{i=1}^\infty\}$  which has measure zero by assumption, and potentially a null set within  $\{(w, \xi) : w \in \mathbb{R}^d, \xi = \xi_i\}$  for  $i \in \{1, 2, \dots\}$ .

For an  $a' \in \mathbb{R}$ , if  $a' \geq a_j$ ,

$$\{(w, \xi) \in \mathbb{R}^{d+p} : \tilde{\nabla}F_j(w, \xi) > a'\} = \bigcup_{i=1}^\infty \{w \in \mathbb{R}^d, \xi = \xi_i : \tilde{\nabla}F_j(w, \xi) > a'\},$$

otherwise

$$\begin{aligned} \{(w, \xi) \in \mathbb{R}^{d+p} : \tilde{\nabla}F_j(w, \xi) > a'\} &= \bigcup_{i=1}^\infty \{w \in \mathbb{R}^d, \xi = \xi_i : \tilde{\nabla}F_j(w, \xi) > a'\} \\ &\cup \{(w, \xi) : w \in \mathbb{R}^d, \xi \in \mathbb{R}^p \setminus (\bigcup_{i=1}^\infty \xi_i)\}, \end{aligned}$$

showing that  $\tilde{\nabla}F(w, \xi)$  is Borel measurable for  $(w, \xi) \in \mathbb{R}^{d+p}$ .  $\blacksquare$

**Proof of Proposition 22** For the goal of showing that functions are definable in the o-minimal structure of  $\mathbb{R}_{\text{exp}, <}$ , we will give a very short background on definable sets, which will serve our proofs. An **atomic formula** is a relation symbol  $\{>, =\}$  applied to **terms** which are made up of finitely many applications of the functions  $\{+, \cdot, \text{exp}\}$  to variables  $\{w_1, w_2, \dots\}$  and constants taken from  $\mathbb{R}$ .<sup>2</sup> **Formulas** are finitely many applications of boolean operations  $\{\vee, \wedge, \neg\}$  and the existential quantifier  $\exists$  to atomic formulas.

2. O-minimality is shown for definable sets with parameters, in our case, taken from  $\mathbb{R}$ , so we can extend the constants from  $\{0, 1\}$  to  $\mathbb{R}$ .

For a formula  $\phi$ , a definable set in  $\mathbb{R}^d$  is the subset  $X \subseteq \mathbb{R}^d$  such that  $\phi$  is true. Let  $\Gamma_g := \{(w, y) \in \mathbb{R}^{d+q} : g(w) = y\}$  be the graph of a function  $g : \mathbb{R}^d \rightarrow \mathbb{R}^q$ . A definable function is a function whose graph is definable, and a function  $g : \mathbb{R}^d \rightarrow \mathbb{R}^q$  is definable if and only if each of its coordinate functions  $g_i(w)$  for  $i \in [q]$  are definable (Coste, 1999, Exercise 1.10). The composition of definable functions is definable, as is the addition and multiplication of definable functions (Coste, 1999, Exercise 1.11).

**Affine Map (AM):** The graph of  $\text{AM}(w, b) := \langle w, x \rangle + b$  for variables  $w \in \mathbb{R}^d$ ,  $b \in \mathbb{R}$ , and constants  $x \in \mathbb{R}^d$ ,

$$\Gamma_{\text{AM}} = \{(w, b, y) \in \mathbb{R}^{d+2} : \langle w, x \rangle + b = y\}$$

is definable, its gradient  $\nabla \text{AM}(w, b) = [x^T, 1]^T$  is measurable as a constant, with  $\nabla \text{AM}(w, b) = \bar{\partial} \text{AM}(w, b)$  since it is continuous.

**ReLU:** The graph of  $\text{ReLU}(w)$  for  $w \in \mathbb{R}$ ,

$$\Gamma_{\text{ReLU}} = \{(w, y) \in \mathbb{R}^2 : (w > 0 \wedge w = y) \vee (\neg(w > 0) \wedge y = 0)\}$$

is definable, its bp derivative  $\text{ReLU}'(w) = \mathbb{1}_{\{w \in \mathbb{R} : w > 0\}}$  is the indicator function of a measurable set, and for  $w \in \mathbb{R}$ ,  $\text{ReLU}'(w) \in \bar{\partial} \text{ReLU}(w)$ , which equals the subdifferential from convex analysis (Clarke, 1990, Proposition 2.2.7).

**Conv2d:** Each of its component functions is an affine mapping, i.e. a filter being applied to the input layer, hence each component function of Conv2d is definable with a measurable gradient.

**MaxPool2d:** Each of its component functions equals  $\text{MP}(w) := \max_{\substack{i \in [H]_{-1} \\ j \in [W]_{-1}}} w_{ij}$  for a subset  $w \in \mathbb{R}^{H \times W}$  of the input layer. The graph of this function can be written as

$$\Gamma_{\text{MP}} = \{(w, y) \in \mathbb{R}^{H \times W + 1} : \bigvee_{\substack{i \in [H]_{-1} \\ j \in [W]_{-1}}} ((w_{ij} = y) \wedge \bigwedge_{\substack{k \in [H]_{-1} \\ l \in [W]_{-1}}} \neg(w_{kl} > w_{ij}))\}.$$

Looping through  $(i, j)$ , the bp gradient of  $\text{MP}(w)$  is set to 1 for the first pair of indices  $(i, j)$  such that  $w_{i,j} = \text{MP}(w)$ , with the remaining entries of the gradient set to 0. The bp gradient can be expressed recursively as

$$\nabla_{ij} \text{MP}(w) = \left( \prod_{\substack{k \in [H]_{-1} \\ l \in [W]_{-1}}} \mathbb{1}_{\{w \in \mathbb{R}^{H \times W} : w_{ij} \geq w_{kl}\}} \right) \left( 1 - \sum_{\substack{k \in [i]_{-1} \\ l \in [W]_{-1}}} \nabla_{kl} \text{MP}(w) - \sum_{l \in [j]_{-1}} \nabla_{il} \text{MP}(w) \right),$$

which is the product and subtraction of real-valued measurable functions. Let  $E(w)$  be the set of pairs  $(i, j)$  such that  $w_{ij} = \text{MP}(w)$ ,  $E(w) := \{(i, j) \in [H]_{-1} \times [W]_{-1} : w_{ij} = \text{MP}(w)\}$ , and let  $e_{ij}$  be a matrix of dimension  $H \times W$  equal to 1 at entry  $(i, j)$  and equal to 0 otherwise, for each  $(i, j) \in E(w)$ . The Clarke subdifferential of  $\text{MP}(w)$  equals  $\bar{\partial} \text{MP}(w) = \text{co}\{e_{ij} : (i, j) \in E(w)\}$  (Clarke, 1990, Proposition 2.3.12), hence  $\nabla \text{MP}(w) \in \bar{\partial} \text{MP}(w)$ .



**Crossentropyloss (CL):** For  $C$  classes, Crossentropyloss takes the form of

$$\text{CL}(w) := -\log \left( \frac{e^{w_t}}{\sum_{i=0}^{C-1} e^{w_i}} \right),$$

where  $t$  is the index of the target class. The graph of this function can be written as

$$\Gamma_{\text{CL}(w)} = \{(w, y) \in \mathbb{R}^{C+1} : \sum_{i=0}^{C-1} e^{w_i} = e^y e^{w_t}\},$$

hence  $\text{CL}(w)$  is definable. The gradient is continuous, with components equal to

$$\nabla_t \text{CL}(w) = \frac{e^{w_t}}{\sum_{i=0}^{C-1} e^{w_i}} - 1$$

and for  $j \neq t$ ,

$$\nabla_j \text{CL}(w) = \frac{e^{w_j}}{\sum_{i=0}^{C-1} e^{w_i}},$$

therefore measurable with  $\nabla \text{CL}(w) = \bar{\partial} \text{CL}(w)$ . ■

## Appendix F. Details of Section 9

DETAILS OF THE NEURAL NETWORK ARCHITECTURE:

Following Pytorch, let  $\text{Conv2d}(i, o, k)$  denote a 2D convolutional layer with  $i$  input and  $o$  output channels, using  $k \times k \times i$  sized filters, with a stride of 1 and 0 padding. Let  $\text{MaxPool2d}(2, 2)$  be a 2D max pool layer with a window size of  $2 \times 2$ , stride of 2, and 0 padding, and let  $\text{Linear}(i, o)$  be a fully connected layer with  $i$  and  $o$  being the number of inputs and outputs. The trained neural network then takes the following form:

$$\begin{aligned} & \text{Input} \rightarrow \text{Conv2d}(1, 6, 5) \rightarrow \text{ReLU} \rightarrow \text{MaxPool2d}(2, 2) \rightarrow \text{Conv2d}(6, 16, 5) \\ & \rightarrow \text{ReLU} \rightarrow \text{MaxPool2d}(2, 2) \rightarrow \text{Conv2d}(16, 120, 4) \rightarrow \text{ReLU} \rightarrow \text{Linear}(120, 84) \\ & \rightarrow \text{ReLU} \rightarrow \text{Linear}(84, 10) \rightarrow \text{CrossEntropyLoss} \rightarrow \text{Output}. \end{aligned}$$

OVERVIEW OF THE BNB IMPLEMENTATION:

We note that since  $\{p_i\} \subset \mathbb{Z}_{>0}$  for this specific application, with  $m = \lfloor (1-s)d \rfloor$ , dynamic programming could have been used to compute  $\Pi_{C \cap \bar{B}_\beta}(\cdot)$ , but the BNB approach was implemented to be applicable for  $\{p_i\} \subset \mathbb{R}_{>0}$ ,  $m \in \mathbb{R}_{>0}$ , and the real-valued objective coefficients of (5), which follows Section 3 where minimal assumptions were placed on  $\{p_i\}$  and  $m$ .

For simplicity let  $y_i := (\|w^i\|_2^2 - \|\max(|w^i| - \beta, 0)\|_2^2)$ . For each node of the search tree, where a 0-1 knapsack problem is considered with subsets of the decision variables already

assigned 1 and 0, the lower bound of the problem uses the Greedy-Split algorithm described in (Kellerer et al., 2004, Chapter 2.1) for the undetermined decision variables. Let  $s$  be the index of the critical item (Kellerer et al., 2004, Section 2.2). The upper bound is computed following (Kellerer et al., 2004, Equation 5.12) when there exists valid indices  $s-1$  and  $s+1$ , or else as (Kellerer et al., 2004, Equation 5.9) when  $s=0$ , where for both computations the floor operators are not employed as generally  $\{y_i\} \notin \mathbb{Z}_{>0}$ .

Before branching if the upper bound is greater than the global lower bound, significant algorithm speed up was observed by checking if  $z_s$  dominates or is dominated by a  $z_i$  already set to 0 or 1. We say that  $z_j$  dominates  $z_k$  if either  $y_j \geq y_k$  and  $p_j < p_k$  or  $y_j > y_k$  and  $p_j \leq p_k$ . Before branching with  $z_s = 0$ , we checked that  $z_s$  does not dominate a  $z_i = 1$ , and before branching with  $z_s = 1$ , we checked that  $z_s$  is not dominated by a  $z_i = 0$ . If one of these cases occurred, the branch was abandoned as the resulting solution would not be optimal.

ESTIMATION OF  $L_0$ ,  $Q$ , AND  $\Delta$ :

We took  $q = 250$  samples  $\{\xi_i\}_{i=1}^q$  and  $q$  pairs of sampled points  $\{(w_j, v_j)\}_{j=1}^q \subset \overline{B}_\kappa$ . The points  $w_j$  are uniformly sampled in  $\overline{B}_{\kappa-\iota}$  for  $\iota = 0.01/\kappa$ , and each point  $v_j$  is sampled uniformly near  $w_j$ , in  $w_j + \overline{B}_\iota$ . The estimate of  $L_0(\xi_i)$  is

$$\widehat{L}_0(\xi_i) = \max_{j \in [q]} \frac{|F(w_j) - F(v_j)|}{\|w_j - v_j\|_2}.$$

The estimate of  $L_0$  is  $\widehat{L}_0 = \text{mean}(\widehat{L}_0(\xi_i))$  and the estimate of  $Q$  equals  $\widehat{Q} = \text{mean}(\widehat{L}_0(\xi_i)^2)$ . For each run, an estimate of  $\Delta$ ,  $\overline{\Delta}_l$  for  $l \in [3]$ , was computed. Given the randomly generated  $w_0^l$ ,  $w_1^l = \Pi_{S \cap \overline{B}_\beta}(w_0^l)$ . For each  $\xi_i$  of the training set,  $F(w_1^l + u_i, \xi_i)$  was computed for a sample  $u_i \sim P_u$ , and the average of these values, over  $i$ , were computed to estimate  $f_\alpha(w_1^l)$ , which was taken as  $\overline{\Delta}_l$  given that  $f_\alpha(w^*) \geq 0$ . The estimate of  $\Delta$  was then chosen as  $\overline{\Delta} = \max_{l \in [3]} \overline{\Delta}_l$ .

## References

- Robert G. Bartle. *The Elements of Integration and Lebesgue Measure*. John Wiley & Sons, 1995.
- Heinz H. Bauschke, D. Russell Luke, Hung M. Phan, and Xianfu Wang. Restricted Normal Cones and Sparsity Optimization with Affine Constraints. *Foundations of Computational Mathematics*, 14(1):63–83, 2014.
- Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 2022. Online first.
- Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021.

- Frank H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.
- Michel Coste. *An introduction to  $\phi$ -minimal geometry*. Institut de Recherche Mathématique de Rennes, 1999.
- Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic Subgradient Method Converges on Tame Functions. *Foundations of Computational Mathematics*, 20(1):119–154, 2020.
- Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, 1999.
- Saeed Ghadimi and Guanghui Lan. Stochastic First- and Zeroth-order Methods for Nonconvex Stochastic Programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.
- A.A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.
- A. M. Gupal. A method for the minimization of almost-differentiable functions. *Cybernetics*, 13(1):115–117, 1977.
- Juha Heinonen. *Lectures on Lipschitz analysis*. University of Jyväskylä, 2004.
- Torsten Hoefer, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021.
- Hans Kellerer, Ulrich Pferschy, and David Pisinger. *Knapsack Problems*. Springer, 2004.
- Guy Kornowski and Ohad Shamir. Oracle Complexity in Nonsmooth Nonconvex Optimization. In *Advances in Neural Information Processing Systems*, volume 34, pages 324–334, 2021.
- Michael R. Metel and Akiko Takeda. Perturbed Iterate SGD for Lipschitz Continuous Loss Functions. *Journal of Optimization Theory and Applications*, 195(2):504–547, 2022.
- Boris S Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Springer, 2013.
- Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.
- Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.
- V. I. Norkin. Generalized-differentiable functions. *Cybernetics*, 16(1):10–12, 1980.
- R. Tyrrell Rockafellar and Roger J-B Wets. *Variational Analysis*. Springer, 2009.
- Andrzej Ruszczyński. Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization. *Optimization Letters*, pages 1–11, 2020.

Steven E. Shreve. *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer, 2004.

Alex Wilkie. O-minimal structures. *Séminaire Bourbaki*, 985:131–142, 2007.

Yi Xu, Rong Jin, and Tianbao Yang. Non-asymptotic Analysis of Stochastic Methods for Non-Smooth Non-Convex Regularized Problems. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of Finding Stationary Points of Nonconvex Nonsmooth Functions. In *International Conference on Machine Learning*, pages 11173–11182, 2020.