

# Asymptotics of constrained $M$ -estimation under convexity

VICTOR-EMMANUEL BRUNEL \*

**Abstract:**  $M$ -estimation, *aka* empirical risk minimization, is at the heart of statistics and machine learning: Classification, regression, location estimation, etc. Asymptotic theory is well understood when the loss satisfies some smoothness assumptions and its derivatives are dominated locally. However, these conditions are typically technical and can be too restrictive or heavy to check. Here, we consider the case of a convex loss function, which may not even be differentiable: We establish an asymptotic theory for  $M$ -estimation with convex loss (which needs not be differentiable) under convex constraints. We show that the asymptotic distributions of the corresponding  $M$ -estimators depend on an interplay between the loss function and the boundary structure of the set of constraints. We extend our results to  $U$ -estimators, building on the asymptotic theory of  $U$ -statistics. Applications of our work include, among other, robust location/scatter estimation, estimation of deepest points relative to depth functions such as Oja’s depth, etc.

*Key words and phrases:* Constrained  $M$ -estimation, empirical risk minimization, convex loss, convex analysis, consistency, asymptotic distribution,  $U$ -statistics, metric projections, directional derivatives..

## 1. INTRODUCTION

### 1.1 Preliminaries

We consider a sequence  $X_1, X_2, \dots$  of independent, identically distributed (iid) random variables taking values in some measurable space  $(E, \mathcal{E})$  and we denote by  $P$  their distribution. Let  $\Theta_0 \subseteq \mathbb{R}^d$  be a non-empty set, which can be interpreted as a parameter space. Here,  $d \geq 1$  is a fixed integer representing the parameter dimension.

Let  $\phi : E \times \Theta_0 \rightarrow \mathbb{R}$  be a function such that  $\phi(\cdot, \theta)$  is measurable and in  $L^1(P)$ , for all  $\theta \in \Theta_0$ . Set  $\Phi(\theta) = \mathbb{E}[\phi(X_1, \theta)]$ , for all  $\theta \in \Theta_0$ . The goal of  $M$ -estimation (or empirical risk minimization) is to estimate a minimizer of  $\Phi$  when only finitely many samples from  $P$  are available. For  $n \geq 1$  and  $\theta \in \Theta_0$ , let  $\Phi_n(\theta) = \frac{1}{n} \sum_{i=1}^n \phi(X_i, \theta)$ . For  $\theta \in \Theta$ ,  $\Phi(\theta)$  is called the population risk evaluated at  $\theta$ , while  $\Phi_n(\theta)$  is the empirical risk based on  $X_1, \dots, X_n$ . The idea of  $M$ -estimation is to use the random function  $\Phi_n$  as a surrogate for  $\Phi$  and estimate a minimizer of  $\Phi$  by selecting a minimizer of  $\Phi_n$ . When minimization is performed over the whole parameter space  $\Theta_0$ , we talk about unconstrained  $M$ -estimation, or simply  $M$ -estimation. If we minimize  $\Phi_n$  on a closed subset  $\Theta$  of  $\Theta_0$ , we talk about constrained  $M$ -estimation with  $\Theta$  as the set of constraints. In this work, we are concerned with the latter.

---

\*CREST-ENSAE, victor.emmanuel.brunel@ensae.fr

Let  $\Theta^* \subseteq \Theta$  be the set of minimizers of  $\Phi$  on  $\Theta$  and assume it is not empty. For all  $n \geq 1$ , let  $\hat{\theta}_n$  be a minimizer of  $\Phi_n$  (provided it exists and can be chosen in a measurable way - see Section 2.2 below). Standard asymptotic theory questions (weak or strong) consistency and aims at determining the asymptotic distribution of a rescaled version of the  $M$ -estimator. That is, does  $d(\hat{\theta}_n, \Theta^*)$  converge (in probability or almost surely) to zero as  $n \rightarrow \infty$ ? Here,  $d(\hat{\theta}_n, \Theta^*)$  is simply the distance of  $\hat{\theta}_n$  to the non-empty set  $\Theta^*$ . If  $\Theta^*$  reduces to a singleton  $\Theta^* = \{\theta^*\}$ , does  $\sqrt{\rho_n}(\hat{\theta}_n - \theta^*)$  converge in distribution for some rescaling factor  $\rho_n \xrightarrow{n \rightarrow \infty} \infty$  and if so, what is the asymptotic distribution?

It may be convenient to consider, instead of  $\hat{\theta}_n$ , a near minimizer of  $\Phi_n$ , that is, a random variable  $\tilde{\theta}_n$  satisfying  $\Phi_n(\tilde{\theta}_n) \leq \inf_{\theta \in \Theta} \Phi_n(\theta) + \varepsilon_n$  where  $\varepsilon_n$  is a (possibly random) small enough error term. For simplicity, here, we only study the properties of exact empirical risk minimizers.

Our main working assumption is that the loss function is convex in its second argument. That is,  $\Theta_0$  and  $\Theta$  are convex sets and  $\phi(x, \cdot)$  is convex on  $\Theta_0$  for  $P$ -almost all  $x \in E$ . Relevant examples include:

1. Location estimation:  $E = \Theta_0 = \mathbb{R}^d$ ,  $\phi(x, \theta) = \ell(x - \theta)$  for some convex function  $\ell : \mathbb{R}^d \rightarrow \mathbb{R}$ . For instance, if  $\ell$  is the squared Euclidean norm, we recover mean estimation. If  $\ell$  is the Euclidean norm, we recover geometric median estimation. If  $\ell(x) = \|x\| - (1 - 2\alpha)u^\top x$ , where  $\alpha \in (0, 1)$  and  $u \in \mathbb{R}^d$  with  $\|u\| = 1$  are fixed ( $\|\cdot\|$  being the Euclidean norm), we recover geometric quantile estimation (e.g., if  $d = 1$  and  $u = 1$ ,  $\Theta^*$  is simply the set of  $\alpha$ -quantiles of  $P$ ). Huber's  $M$ -estimators, adding robustness to mean estimators, correspond to the loss  $\ell(x) = h_c(\|x\|)$ ,  $x \in \mathbb{R}^d$ , where for all  $t \geq 0$ ,  $h_c(t) = t^2$  if  $t \leq c$ ,  $h_c(t) = 2ct - c^2$  if  $t > c$  and  $c > 0$  is a given, tuning parameter.
2. Location estimation on matrix spaces: Let  $E = \Theta_0 =: \mathcal{S}_d^+$  be the space of  $d \times d$  symmetric, positive semi-definite matrices. There are several ways of averaging positive definite matrices, beyond simply taking their arithmetic mean (i.e., their standard linear average). A simple example is that of the harmonic mean, which is simply the inverse of the linear average of the inverses (if the matrices are positive definite). More involved ways include (again for positive definite matrices) the Karsher mean, which, in the case of 2 such matrices, reduces to their geometric mean [7]. In the context of optimal transport, a large body of literature has been interested in the Bures-Wasserstein mean of positive definite matrices, which is related to Wasserstein barycenters on the set of Gaussian distributions [2, 54]. In fact, it is shown in [30, Lemma A.5] that the Bures-Wasserstein mean is the solution to a convex optimization problem. Hence, as it is done in [30], the Bures-Wasserstein barycenter of iid, random, positive (semi-)definite matrices can be analyzed under the prism of  $M$ -estimation with convex loss, and our results also allows to consider the constrained case, as well as robust alternatives to Bures-Wasserstein barycenters (such as the Bures-Wasserstein median, see [2]).
3. Linear regression (here, data are rather denoted as pairs  $(X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ ,  $n \geq 1$ ):  $E = \mathbb{R}^d \times \mathbb{R}$ ,  $\Theta = \mathbb{R}^d$ ,  $\phi((x, y), \theta) = \ell(y - \theta^\top x)$  for some  $\ell : \mathbb{R} \rightarrow \mathbb{R}$  (which, again in our context, we assume to be convex). If  $\ell(t) = t^2$ , we recover least squares estimation. If  $\ell(t) = |t|$ , this is median regression, etc.

In all these examples, we can take  $\Theta_0 = \Theta = \mathbb{R}^d$  (or  $\mathcal{S}_d^+$ ), corresponding to unconstrained estimation, but we could also assume that  $\Theta$  is a closed, strict subset of  $\Theta_0$ . Perhaps the simplest example is the case when  $E = \Theta_0 = \mathbb{R}^d$ ,  $\Theta \subseteq \mathbb{R}^d$  is a compact convex subset and  $\phi(x, \theta) = \|x - \theta\|^2$ . In that case, it is easy to check that  $\theta^* = \pi_\Theta(\mathbb{E}[X])$  and  $\hat{\theta}_n = \pi_\Theta(\bar{X}_n)$  are the unique minimizers of  $\Phi$  and  $\Phi_n$  respectively, where  $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$  and  $\pi_\Theta$  is the metric projection on  $\Theta$ . Of course, this example can be studied with elementary tools, but it is worth keeping it in mind as an illustration of our results, in order to fix ideas.

Typically, proving consistency and finding the asymptotic distribution of  $M$ -estimators require some tools from the theory of empirical processes and imposes some smoothness of the loss function  $\phi$  in its second argument. Moreover, it is often assumed that the partial derivatives of  $\phi$ , with respect to its second argument, are locally dominated, allowing the use of dominated convergence to swap derivatives and expectations in the analysis. In our context, the full power of convexity comes in through fairly elementary convex analysis and allows to completely avoid such common technical assumptions.

## 1.2 Related works

$M$ -estimation is a quintessential problem in statistical inference (maximum likelihood estimation being a particular instance in general) and, as a particular case, constrained  $M$ -estimation.

Asymptotic theory of statistical estimation has been overlooked in the era of high-dimensional data and models. Yet, it provides benchmarks for non-asymptotic theory and asymptotic approximations produce less conservative inference than non-asymptotic approaches, and they are relevant when the data set contains a lot of samples and their dimension is not too large.

Asymptotic theory of  $M$ -estimators is well understood when the loss function is smooth and satisfies local domination properties [31, 55, 56]. Under similar smoothness and domination assumptions, [18] also derived asymptotic properties in the constrained case, when the set of constraints is a regular closed set and the population minimizer is a local minimum of the population risk in the ambient space. See also [34] for inference on constrained statistical problems and [26, 47] for special cases. Recently, [35] drew connections between the statistical error of constrained  $M$ -estimation and the statistical dimension of the constrained set, building on [11, 46] in linear regression and Gaussian sequence models. Even though these connections belong to the non-asymptotic world, we also discuss such connections at infinitesimal scales in the remarks following Theorem 7 below.

When the loss function is convex, [19] proved asymptotic normality, only requiring the population risk (that is,  $\Phi$ ) being twice differentiable at the (unique) population minimizer, with positive definite Hessian at that point - convexity allowing to avoid any local domination assumption. [40] proved further asymptotic expansions of the statistical error under stronger smoothness assumptions of convex the loss.

Asymptotics of penalized  $M$ -estimators have also been established [24], in particular for penalized regression (such as Lasso) [27].

In the context of high dimensional linear regression and classification, some recent work has also tackled the asymptotics of penalized  $M$ -estimators and bagged penalized  $M$  estimators in growing dimension (that is, when the dimension  $d$  also diverges with the sample size) [5, 6, 29]. Related to this line of work are the high-dimensional central limit theorems of [12, 15] which correspond to the squared Euclidean loss in the context of  $M$ -estimation. To the best of our knowledge, similar high-dimensional central limit theorems have not been tackled for general  $M$ -estimators.

This work is not concerned with penalized  $M$ -estimation. Indeed, even though penalized and constrained optimization problems are related through Lagrangian functions, in penalized statistical problems, it is standard to let the penalty depend on the sample size in order to enforce some regularization and achieve optimal performance, although here, we only consider fixed constraint sets, independently of the sample size.

## 1.3 Outline

In Section 2, we give some key lemmas that we use in our main results. Section 2.1 gathers some results about convex functions and sequences of convex functions, which we chose to highlight in the first part of this work because they are essential to build the intuition behind the theory. In Section 2.2, which is much more theoretical and could be skipped at first, we deal with the

existence of a measurable empirical minimizer, based on results that guarantee the existence of measurable selections. Section 3 focuses on consistency of convex  $M$ -estimators and Section 4 deals with asymptotic distributions of  $M$ -estimators. We propose an extension to  $U$ -estimators with convex loss in Section 5. More lemmas about convex functions, convex sets and cones, and metric projections, which are only used for some technical parts of the main proofs, but not essential to build the intuition, are deferred to the appendix. However, Section C, in the appendix, on directional differentiability of metric projections onto convex sets, may be of independent interest to the reader.

## 1.4 Notation and standard definitions/assumptions

Here, we gather all the notation that we use in this work, as well as several simple definitions.

1. In this work,  $(\Omega, \mathcal{F}, \mathbb{P})$  is a fixed probability space and we assume that all the random variables that we consider are defined on that space. We let  $X_1, X_2, \dots$  be iid random variables with values in a measurable space  $E$  and we let  $P = X_1 \# \mathbb{P}$  be their distribution. The set  $\Theta_0$  is a fixed, open, convex subset of  $\mathbb{R}^d$  and  $\Theta$  is a closed, convex subset of  $\Theta_0$ . The loss function  $\phi : E \times \Theta_0 \rightarrow \mathbb{R}$  is assumed to be measurable in its first argument and convex in its second, and to satisfy  $\phi(\cdot, \theta) \in L^1(P)$  for all  $\theta \in \Theta_0$ . We let  $\Phi(\theta) = \mathbb{E}[\phi(X_1, \theta)]$  for all  $\theta \in \Theta_0$  (referred to as *population risk*) and for all  $n \geq 1$ ,  $\omega \in \Omega$  and  $\theta \in \Theta_0$ ,  $\Phi_n(\omega, \theta) = n^{-1} \sum_{i=1}^n \phi(X_i(\omega), \theta)$  (referred to as *empirical risk*). For simplicity, unless this amount of precision is needed, we simply write  $\Phi_n(\theta)$  and skip the dependence on  $\omega \in \Omega$ .
2. The power set of a non-empty set  $A$  is denoted by  $\mathcal{P}(A)$ .
3. Given a subset  $G \subseteq \mathbb{R}^d$ , we denote by  $\text{int}(G)$  its interior,  $\text{cl}(G)$  its closure and  $\partial G = \text{cl}(G) \setminus \text{int}(G)$  its boundary.
4. Any symmetric, positive definite matrix  $S \in \mathbb{R}^{d \times d}$  yields a scalar product by setting, for  $x, y \in \mathbb{R}^d$ ,  $\langle x, y \rangle_S := x^\top S y$ . The associated Euclidean norm is given by  $\|x\|_S = \langle x, x \rangle_S^{1/2}$  for all  $x \in \mathbb{R}^d$ . The corresponding Euclidean ball with center  $x \in \mathbb{R}^d$  and radius  $r \geq 0$  is denoted by  $B_S(x, r)$ .
5. Given a vector  $u \in \mathbb{R}^d$ , the linear subspace of  $\mathbb{R}^d$  that is orthogonal to  $u$  with respect to  $\langle \cdot, \cdot \rangle_S$  is denoted by  $u^\perp_S$ : If  $u = 0$ ,  $u^\perp_S = \mathbb{R}^d$  and if  $u \neq 0$ ,  $u^\perp_S$  is some linear hyperplane. When  $L \subseteq \mathbb{R}^d$ , we denote by  $L^\perp_S$  the linear subspace of  $\mathbb{R}^d$  that is orthogonal to  $L$  with respect to  $\langle \cdot, \cdot \rangle_S$ .
6. For a set  $C \subseteq \mathbb{R}^d$ , a vector  $u \in \mathbb{R}^d$  and a real number  $t \in \mathbb{R}$ , we denote by  $C_{u,t}^S = \{x \in C : \langle u, x \rangle_S = t\}$ , which may be empty. When  $t = 0$ , we simply write  $C_u^S = C_{u,0}^S$ .
7. The distance of a point  $x \in \mathbb{R}^d$  to a closed set  $C \subseteq \mathbb{R}^d$  with respect to the Euclidean norm associated with  $S$  is denoted by  $d_S(x, C) = \min_{y \in C} \|x - y\|_S$ .
8. The metric projection onto a non-empty, closed convex set  $C \subseteq \mathbb{R}^d$  with respect to  $\langle \cdot, \cdot \rangle_S$  is denoted by  $\pi_C^S$ : For all  $u \in \mathbb{R}^d$ ,  $\pi_C^S(u)$  is the unique minimizer of the map  $t \in C \mapsto \|t - u\|_S^2$ . In particular,  $d_S(u, C) = \|u - \pi_C^S(u)\|_S$ .
9. Let  $G \subseteq \mathbb{R}^d$  be a non-empty, closed, convex set and  $x_0 \in G$ . The tangent cone to  $G$  at  $x_0$  is the set of all  $t \in \mathbb{R}^d$  such that  $x_0 + \varepsilon t \in G$  for all small enough  $\varepsilon > 0$ . It is a convex cone, not necessarily closed. Its closure is called the support cone to  $G$  at  $x_0$ . Let  $S \in \mathbb{R}^{d \times d}$  be symmetric, positive definite. The normal cone to  $G$  at  $x_0$  with respect to  $S$  is the set of all  $t \in \mathbb{R}^d$  satisfying  $\langle t, x - x_0 \rangle_S \leq 0$  for all  $x \in G$ . It is a closed, convex cone. When there is no mention of a matrix  $S$ , it is implicitly assumed to be the identity matrix.
10. The support function of a non-empty convex set  $C \subseteq \mathbb{R}^d$  is the map  $h_C : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  defined by  $h_C(t) = \sup_{u \in C} u^\top t$ . If  $t \neq 0$ , it is the largest (signed) distance from the origin to a hyperplane orthogonal to  $t$  and that is tangent to  $C$ . It is easy to check that  $h_C$  is a sublinear function (that is, positively homogeneous and convex). If  $C$  is bounded, then  $h_C$  only takes finite values. See, e.g., [49, Section 1.7.1].

11. In all notation above, when  $S$  is the identity matrix, we drop the subscript or superscript  $S$  and simply write, for instance,  $\|x\|$ ,  $B(x, r)$ ,  $u^\perp$ ,  $C_u$ ,  $\pi_C$ , etc.
12. Given a set  $C \subseteq \mathbb{R}^d$  and a function  $f : C \rightarrow \mathbb{R}$ , the set of minimizers (resp. maximizers) of  $f$  on  $C$  is denoted by  $\text{Argmin}_{y \in C} f(y)$  (resp.  $\text{Argmax}_{y \in C} f(y)$ ). This set may be empty. When this set is a singleton, we denote by  $\text{argmin}_{y \in C} f(y)$  (resp.  $\text{argmax}_{y \in C} f(y)$ ), with lower case “a”, the unique element of that set.
13. Let  $f$  be a function defined on a subset of  $\mathbb{R}^d$ , with values in  $\mathbb{R}^p$  for some  $p \geq 1$  (for us, in practice,  $p = 1$  or  $d$ ). Then, given a point  $x$  in the interior of the domain of  $f$ , we say that  $f$  has a directional derivative at  $x$  in the direction  $t \in \mathbb{R}^d$  if and only if the quantity  $\varepsilon^{-1}(f(x + \varepsilon t) - f(x))$  has a limit as  $\varepsilon \rightarrow 0$ , with  $\varepsilon > 0$ . In that case, we denote this limit by  $d^+f(x; t)$ . Note that if  $f$  has directional derivatives at  $x \in \mathbb{R}^d$ , then it must be continuous at  $x$ . Moreover, the map  $d^+f(x; \cdot)$  is automatically measurable, since the limit can be taken along the sequence  $\varepsilon = 1/k$ ,  $k \geq 1$ . If the ratio  $\varepsilon^{-1}(f(x + \varepsilon t) - f(x))$  converges uniformly in  $t$  on all compact subsets of  $\mathbb{R}^d$ , we say that  $f$  has directional derivatives at  $x$  in Hadamard sense. This is equivalent to requiring that for all  $t \in \mathbb{R}^d$ , for all sequences  $(t_n)_{n \geq 1}$  converging to  $t$  and for all sequences  $(\varepsilon_n)_{n \geq 1}$  of positive numbers converging to 0,  $\varepsilon_n^{-1}(f(x + \varepsilon_n t_n) - f(x))$  has a (finite) limit as  $n \rightarrow \infty$  (see, e.g., [17, Chapter III]).
14. If  $f$  is differentiable at  $x$ , we denote by  $df(x; \cdot)$  its differential. That is,  $df(x; t) = d^+f(x, t) = \nabla f(x)^\top t$  for all  $t \in \mathbb{R}^d$ .
15. Given a convex set  $G_0 \subseteq \mathbb{R}^d$ , when we talk about a convex function on  $G_0$ , we always mean that it takes finite values only, i.e., we only consider convex functions  $f : G_0 \rightarrow \mathbb{R}$ , which may be the restriction to  $G$  of some lower-semicontinuous convex function  $\tilde{f} : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  whose domain contains  $G_0$ .
16. We call *random convex function* any map  $f : \Omega \times G \rightarrow \mathbb{R}$ , where  $G \subseteq \mathbb{R}^d$  is some convex set, such that  $f(\cdot, t)$  is measurable for all  $t \in G$  and  $f(\omega, \cdot)$  is convex for all  $\omega \in \Omega$ . We could only assume that  $f(\omega, \cdot)$  is convex for  $\mathbb{P}$ -almost all  $\omega \in \Omega$ , but this does not bring significantly more generality. Unless we need to emphasize the dependence on  $\omega$  explicitly, we rather write  $f(t)$  instead of  $f(\omega, t)$  for simplicity.
17. The covariance matrix of a random vector  $X$  in  $\mathbb{R}^d$  with two moments is defined as  $\text{var}(X) = \mathbb{E}[XX^\top] - \mathbb{E}[X]\mathbb{E}[X]^\top = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top]$ . That is, for all vectors  $u, v \in \mathbb{R}^d$ ,  $u^\top \text{var}(X)v = \text{cov}(u^\top X, v^\top X)$ . When  $S \in \mathbb{R}^{d \times d}$  is symmetric, positive definite, we denote by  $\text{var}_S(X) = S \text{var}(X) S = \text{var}(SX)$  so that for all vectors  $u, v \in \mathbb{R}^d$ , we have the identity  $u^\top \text{var}_S(X)v = \text{cov}(\langle u, X \rangle_S, \langle v, X \rangle_S)$ . This is the matrix representation of the covariance operator of  $X$  corresponding to the Euclidean structure defined by  $S$ .
18. For all vectors  $u \in \mathbb{R}^d$  and symmetric, positive semi-definite matrices  $V \in \mathbb{R}^{d \times d}$ , we denote by  $\mathcal{N}_d(u, V)$  the  $d$ -variate Gaussian distribution with mean  $u$  and covariance matrix  $V$ .

## 2. KEY LEMMAS ABOUT DETERMINISTIC AND RANDOM CONVEX FUNCTIONS

### 2.1 On the behavior of convex functions and sequences of convex functions

First, we state a minimum principle for convex functions, which we will use a few times in the next sections.

LEMMA 1. *Let  $G_0 \subset \mathbb{R}^d$  be an open convex set and  $G \subseteq G_0$  be a closed convex subset. Let  $f : G_0 \rightarrow \mathbb{R}$  be a convex function and  $K \subseteq G_0$  be any compact, convex set. If  $\min_{t \in \partial K \cap G} f(t) > f(t_0)$  for some  $t_0 \in K \cap G$ , then  $\text{Argmin}_{t \in G} f(t) \subseteq K$  and it is not empty.*

REMARK 1. • Recall that a convex function defined on an open convex set is automatically

continuous on that set [48, Theorem 10.1], hence, it automatically reaches its bounds on any compact set.

- The phrasing of this lemma is a bit technical, but a simpler version, when  $G = G_0 = \mathbb{R}^d$ , says that if  $f$  has one value inside  $K$  that is smaller than all values taken on  $\partial K$ , then, it has at least one minimizer, and they all lie in  $K$ . We need this slightly more technical statement in order to deal with constrained  $M$ -estimation later.

PROOF. Fix some arbitrary  $t \in G \setminus K$  and let us show that necessarily,  $f(t) > f(t_0)$ . Set  $\phi : \lambda \in [0, 1] \mapsto f(t_0 + \lambda(t - t_0))$ , which is a convex function. First, note that  $t_0 \notin \partial K$  (or else,  $t_0$  would be in  $\partial K \cap G$  so  $f(t_0) \geq \min_{\partial K \cap G} f$ , which would contradict the assumption). Hence, there must be some  $\lambda^* \in (0, 1)$  such that  $t_0 + \lambda^*(t - t_0) \in \partial K$ . Moreover, since both  $t_0$  and  $t$  are in  $G$ ,  $t_0 + \lambda^*(t - t_0) \in G$ . Therefore, by assumption,  $\phi(\lambda^*) > \phi(0)$ . Hence, convexity of  $\phi$  implies that it must be increasing on  $[\lambda^*, 1]$ , yielding that  $\phi(1) \geq \phi(\lambda^*)$  and hence, that  $\phi(1) > \phi(0)$ . That is,  $f(t) > f(t_0)$ .

Therefore, the minimizers (if any) of  $f$  on  $G$  must be contained in  $K$ . Finally, there must be at least one such minimizer since  $f$  is continuous on the compact set  $K \cap G$ .  $\square$

In the main statistical results presented in the next sections, Lemma 1 will be used to localize empirical minimizers of  $\Phi_n$ .

The second key result is due to Rockafellar and shows that, for sequences of convex functions, uniform convergence can be deduced from pointwise convergence on a dense subset. From this lemma, we will derive two probabilistic corollaries.

LEMMA 2. [48, Theorem 10.8] Let  $G_0 \subseteq \mathbb{R}^d$  be an open convex set and  $f, f_1, f_2, \dots$  be convex functions on  $G_0$ . Assume that there is a dense subset  $C$  of  $G_0$  such that for all  $t \in C$ ,  $f_n(t) \rightarrow f(t)$ . Then,  $f_n$  converges uniformly to  $f$  on all compact subsets of  $G_0$ .

An important consequence that we will use extensively is the following corollary.

COROLLARY 1. Let  $f, f_1, f_2, \dots$  be random convex functions defined on an open convex set  $G_0 \subseteq \mathbb{R}^d$ . Assume that  $f_n(t) \xrightarrow[n \rightarrow \infty]{} f(t)$  almost surely (resp. in probability) for all  $t \in G_0$ . Then, for all compact sets  $K \subseteq G_0$ ,  $\sup_K |f_n - f| \xrightarrow[n \rightarrow \infty]{} 0$  almost surely (resp. in probability).

PROOF. Let us prove the statement for the almost sure convergence and the convergence in probability separately.

Almost sure convergence.

Let  $C$  be a dense and countable subset of  $G_0$ . By assumption, for each  $t \in C$ , it holds with probability one that  $f_n(t) \xrightarrow[n \rightarrow \infty]{} f(t)$ . Since  $C$  is countable, this implies that with probability 1,  $f_n(t) \xrightarrow[n \rightarrow \infty]{} f(t)$  for all  $t \in C$  simultaneously. Hence, by Lemma 2, with probability 1,  $f_n$  converges uniformly to  $f$  on all compact subsets of  $G_0$ .

Convergence in probability.

Again, let  $C$  be a dense and countable subset of  $G_0$  and fix a compact subset  $K$  of  $G_0$ . Our goal is to show that  $Z_n := \sup_{t \in K} |f_n(t) - f(t)| \xrightarrow[n \rightarrow \infty]{} 0$  in probability. It is necessary and sufficient to show that every subsequence of  $(Z_n)_{n \geq 1}$  has a further subsequence that converges to 0 almost surely [13, Section 3.3, Lemma 2]. With no loss of generality (since we could just renumber the terms of the sequence), let us prove that  $(Z_n)_{n \geq 1}$  has a subsequence that converges to 0 almost surely. Denote by  $t_1, t_2, \dots$  the elements of  $C$ .



By assumption,  $f_n(t_1) \xrightarrow[n \rightarrow \infty]{} f(t_1)$  in probability, so it has a subsequence that converges almost surely. That is, there is an increasing map  $\psi_1 : \mathbb{N}^* \rightarrow \mathbb{N}^*$  such that  $f_{\psi_1(n)}(t_1) \xrightarrow[n \rightarrow \infty]{} f(t_1)$  almost surely.

Similarly,  $(f_{\psi_1(n)}(t_2))_{n \geq 1}$  being a subsequence of  $(f_n(t_2))_{n \geq 1}$ , it converges almost surely to  $f(t_2)$  and thus has a further subsequence  $(f_{\psi_1(\psi_2(n))}(t_2))_{n \geq 1}$  that converges almost surely to  $f(t_2)$ . By induction, one can construct a sequence of increasing maps  $\psi_p : \mathbb{N}^* \rightarrow \mathbb{N}^*$ ,  $p \geq 1$ , such that for all integers  $p \geq 1$ ,  $f_{\psi_1 \circ \dots \circ \psi_p(n)}(t_p)$  converges to  $f(t_p)$  almost surely. Let  $\psi(n) = \psi_1 \circ \dots \circ \psi_n(n)$ , for all  $n \geq 1$ . This is an increasing map; Let us prove that  $Z_{\psi(n)} \xrightarrow[n \rightarrow \infty]{} 0$  almost surely, which will prove the lemma.

First, note that with probability 1,  $f_{\psi_1 \circ \dots \circ \psi_p(n)}(t_p)$  converges to  $f(t_p)$  simultaneously for all  $p \geq 1$ . Second, for all  $p \geq 1$ ,  $(f_{\psi(n)}(t_p))_{n \geq 1}$  is a subsequence of  $(f_{\psi_1 \circ \dots \circ \psi_p(n)}(t_p))_{n \geq 1}$  (except maybe for the first  $p$  terms of the sequence). Hence,  $f_{\psi(n)}(t_p) \xrightarrow[n \rightarrow \infty]{} f(t_p)$  for all  $p \geq 1$ , almost surely. The rest follows from the first part of the proof (the case of almost sure convergence).  $\square$

In fact, we can also derive a similar corollary for  $L^p$  convergence, for any  $p \geq 1$ . We defer it to the appendix (Section E), because we only use it to formulate an open question, see the end of Section 4.2).

## 2.2 On the existence of measurable minimizers and measurable subgradients

The existence of minimizers of a random convex function can often be established quite easily (for instance, if the function is coercive). Same for subgradients since any convex function defined on an open convex set has at least one subgradient at any point of that set. However, the existence of a measurable minimizer or subgradient is much less trivial and relies on the theory of measurable selections.

### 2.2.1 Measurable selections

**DEFINITION 1.** Let  $\Gamma : \Omega \rightarrow \mathcal{P}(\mathbb{R}^d)$  be a multifunction, that is, a function that maps any  $\omega \in \Omega$  to some non-empty set  $\Gamma(\omega) \subseteq \mathbb{R}^d$ . A measurable selection of  $\Gamma$  is a measurable map  $\gamma : \Omega \rightarrow \mathbb{R}^d$  such that for all  $\omega \in \Omega$ ,  $\gamma(\omega) \in \Gamma(\omega)$ .

There are numerous theorems that guarantee the existence of measurable selections in various setups, see [21, 38]. The one that we will need is the following, that follows from combining Theorems 3.2 (ii), 3.5 and 5.1 of [21]. Denote by  $\mathcal{C}$  the collection of all non-empty, closed subsets of  $\mathbb{R}^d$ .

**LEMMA 3.** Let  $\Gamma : \Omega \rightarrow \mathcal{C}$  be a multifunction. Assume that for all compact sets  $K \subseteq \mathbb{R}^d$ , the set  $\{\omega \in \Omega : \Gamma(\omega) \cap K \neq \emptyset\}$  is measurable (that is, it belongs to the  $\sigma$ -algebra  $\mathcal{F}$ ). Then,  $\Gamma$  has a measurable selection.

A multifunction satisfying this property above is called  $C$ -measurable ( $C$  as in “compact”, the test sets  $K$  used in Lemma 3 being compact).

### 2.2.2 Measurable empirical risk minimizers

From Lemma 3, we obtain the following result, which will guarantee the existence of a measurable empirical risk minimizer for large enough  $n$ , and which will, at the same time, yield its strong consistency.

**THEOREM 1.** Let  $f, f_1, f_2, \dots$  be random convex functions defined on an open convex set  $G_0 \subseteq \mathbb{R}^d$  such that for all  $t \in G_0$ ,  $f_n(t) \xrightarrow[n \rightarrow \infty]{} f(t)$  almost surely. Let  $G \subseteq G_0$  be a closed, convex set. Assume

that  $G^* := \operatorname{Argmin}_{t \in G} f(t)$  is non-empty and compact. Then, there exists a sequence  $(t_n)_{n \geq 1}$  of random variables with values in  $G$  such that with probability 1,  $t_n$  is a minimizer of  $f_n$  on  $G$  for all large enough  $n$ . Moreover,  $d(t_n, G^*) \xrightarrow[n \rightarrow \infty]{} 0$  almost surely.

PROOF. For  $n \geq 1$ , let  $M_n := \operatorname{Argmin}_{t \in G} f_n(t)$ , possibly empty. We proceed in two steps. First, we prove that with probability 1,  $M_n$  is non-empty for all large enough  $n$ . Second, we use the measurable selection to obtain such a sequence  $(t_n)_{n \geq 1}$ .

Step 1. Note that if  $G$  is compact, then  $M_n \neq \emptyset$  for all  $n \geq 1$ , since  $f_n$  is convex, hence continuous, on the open set  $G_0$ .

First, Corollary 1 yields that  $f_n$  converges uniformly to  $f$  on any compact subset of  $G_0$ , almost surely. Fix some arbitrary, small enough  $\varepsilon > 0$  such that  $G_\varepsilon^* := \{t \in \mathbb{R}^d : d(t, G^*) \leq \varepsilon\}$ . This set is compact, so

$$(1) \quad \sup_{t \in G_\varepsilon^* \cap G} |f_n(t) - f(t)| \xrightarrow[n \rightarrow \infty]{} 0.$$

Let  $f^* := \min_{t \in G} f(t)$  be the smallest value of  $f$  on  $G$  (note that  $f^*$  is measurable, since it can be written as the infimum of  $f(t)$  for  $t$  ranging in a countable, dense subset of  $G$ ). Convexity of  $f$  on the open set  $G_0$  implies its continuity. Therefore,  $\eta := \min_{t \in \partial G_\varepsilon^* \cap G} f(t) - f^* > 0$ .

Then, the following holds with probability 1: For all sufficiently large integers  $n$  and for all  $t \in \partial G_\varepsilon^* \cap G$ ,

$$\begin{aligned} f_n(t) &\geq f(t) - \eta/3 && \text{by (1)} \\ &\geq f^* + \eta - \eta/3 && \text{by definition of } \eta \\ &\geq f_n(t^*) - \eta/3 + \eta - \eta/3 && \text{again by (1)} \\ &= f_n(t^*) + \eta/3 > f_n(t^*). \end{aligned}$$

Therefore, by Lemma 1, it holds with probability 1 that, for all large enough integers  $n \geq 1$ ,

$$(2) \quad \emptyset \neq M_n \subseteq G_\varepsilon^*.$$

Step 2. Now, fix an arbitrary element  $t_0 \in G$ . For all integers  $n \geq 1$ , let  $\Gamma_n := \begin{cases} M_n & \text{if } M_n \neq \emptyset \\ \{t_0\} & \text{otherwise.} \end{cases}$

Let us prove that  $\Gamma_n$  has a measurable selection, for all  $n \geq 1$ . Since  $M_n$  is always closed (by continuity of  $f_n$ ),  $\Gamma_n$  is always non-empty and closed, so by Lemma 3, it is sufficient to check that for each  $n \geq 1$ , the multiset function  $\Gamma_n : \Omega \rightarrow \mathcal{C}$  is  $C$ -measurable in order to guarantee the existence of a measurable selection.

Fix  $n \geq 1$  and let  $K \subseteq \mathbb{R}^d$  be any compact set and let us show that the set  $\{\omega \in \Omega : \Gamma_n(\omega) \cap K \neq \emptyset\}$  is a measurable set.

First, rewrite  $\{\omega \in \Omega : \Gamma_n(\omega) \cap K \neq \emptyset\} = \{\omega \in \Omega : M_n(\omega) \cap K \neq \emptyset\} \cup \{\omega \in \Omega : M_n(\omega) = \emptyset, t_0 \in K\}$ . Since  $f_n(\omega, \cdot)$ <sup>1</sup> is continuous for every  $\omega \in \Omega$ , the first set in this union can be rewritten as  $\{\omega \in \Omega : \inf_{t \in G} f_n(\omega, t) = \inf_{t \in K \cap G} f_n(\omega, t)\}$ . Again, using continuity of  $f_n(\omega, \cdot)$  for all  $\omega \in \Omega$ , we can rewrite  $\inf_{t \in G} f_n(\omega, t)$  and  $\inf_{t \in K \cap G} f_n(\omega, t)$  as  $\inf_{t \in \tilde{G}_1} f_n(\omega, t)$  and  $\inf_{t \in \tilde{G}_2} f_n(\omega, t)$  respectively, where  $G_1$  and  $G_2$  are dense, countable subsets of  $G$  and  $K \cap G$  respectively. Therefore, both  $\inf_{t \in G} f_n(\omega, t)$  and  $\inf_{t \in K \cap G} f_n(\omega, t)$  are measurable (as maps from  $\Omega$  to  $\mathbb{R} \cup \{-\infty\}$ ) and we obtain that  $\{\omega \in \Omega : M_n(\omega) \cap K \neq \emptyset\} \in \mathcal{F}$ .

<sup>1</sup>recall that above, we only wrote  $f_n(t)$  instead of  $f_n(\omega, t)$  for simplicity.



Now,  $\{\omega \in \Omega : M_n(\omega) = \emptyset, t_0 \in K\}$  is empty if  $t_0 \notin K$ , which is measurable. If  $t_0 \in K$ , it reduces to the set  $\{\omega \in \Omega : M_n(\omega) = \emptyset\}$ , which can be decomposed as

$$\{\omega \in \Omega : M_n(\omega) = \emptyset\} = \bigcap_{p \in \mathbb{N}^*} \bigcup_{q \geq p+1} \{\omega \in \Omega : \min_{t \in G \cap B(t_0, q)} f_n(\omega, t) < \min_{t \in G \cap B(t_0, q)} f_n(\omega, t)\}$$

which, therefore, is also measurable.

Finally, Lemma 3 implies the existence of a sequence  $(t_n)_{n \geq 1}$  of random variables such that for all  $n \geq 1$ ,  $t_n \in \Gamma_n$ . Furthermore, by Step 1 of this proof, we also obtain that with probability 1,  $t_n \in M_n$  for all large enough  $n$ .

*Step 3.* Finally, following the reasoning of Step 1, (2) yields that for all  $\varepsilon > 0$ , it holds, with probability 1, that  $d(t_n, G^*) \leq \varepsilon$  for all large enough  $n$ . That is,  $d(t_n, G^*) \xrightarrow[n \rightarrow \infty]{} 0$  almost surely.  $\square$

### 2.2.3 Measurable subgradients

Now, we apply Lemma 3 to show the existence of measurable subgradients for random convex functions. Recall that for a convex function  $f$  defined on a convex set  $G_0 \subseteq \mathbb{R}^d$ , a subgradient of  $f$  at a point  $t_0 \in G_0$  is any vector  $u \in \mathbb{R}^d$  such that

$$f(t) \geq f(t_0) + u^\top(t - t_0), \quad \forall t \in G_0.$$

We denote by  $\partial f(t_0)$  the collection of all subgradients of  $f$  at  $t_0$ . If  $t_0 \in \text{int}(G_0)$ , then  $\partial f(t_0)$  is non-empty, compact and convex by Lemma 5. In particular, if  $G_0$  is open, then  $f$  has subgradients at every point of  $G_0$ . Now, if  $f$  is a random convex function, the existence of a measurable subgradient (i.e., that is chosen in a measurable way) at  $t_0 \in \text{int}(G_0)$  is granted by the following theorem.

**THEOREM 2.** *Let  $f$  be a random convex function defined on a convex set  $G_0 \subseteq \mathbb{R}^d$  and let  $t_0 \in \text{int}(G_0)$ . Then,  $f$  has a measurable subgradient at  $t_0$ .*

**PROOF.** Let  $\Gamma = \partial f(t_0)$  be the set of subgradients of  $f$  at  $t_0$  (that is, for all  $\omega \in \Omega$ ,  $\Gamma(\omega) = \partial(f(\omega, \cdot))(t_0)$ ). Since  $t_0 \in \text{int}(G_0)$ ,  $\Gamma$  only takes non-empty values. Moreover, by Lemma 5, it always takes closed values, so  $\Gamma$  is a  $\mathcal{C}$ -valued multifunction. Hence, it is sufficient to check that it is  $\mathcal{C}$ -measurable in order to apply Lemma 3.

Let  $K \subseteq \mathbb{R}^d$  be any arbitrary compact set. Lemma 4 yields that  $\Gamma \cap K \neq \emptyset$  if and only if there exists  $u \in K$  with the property that  $\sup_{t \in B(t_0, \varepsilon)} (u^\top(t - t_0) - f(t) + f(t_0)) \leq 0$  where  $\varepsilon > 0$  is any small enough positive number satisfying that  $B(t_0, \varepsilon) \subseteq \text{int}(G_0)$ . Since  $f$  is convex, it is continuous on  $\text{int}(G)$  and, hence, on  $B(t_0, \varepsilon)$ . Let  $C$  be a fixed dense, countable subset of  $B(t_0, \varepsilon)$ . Then,  $\Gamma \cap K \neq \emptyset$  if and only if there exists  $u \in K$  for which  $\sup_{t \in C} (u^\top(t - t_0) - f(t) + f(t_0)) \leq 0$ . Let  $h(\omega, u) = \sup_{t \in C} (u^\top(t - t_0) - f(\omega, t) + f(\omega, t_0))$ , for all  $\omega \in \Omega$  and  $u \in \mathbb{R}^d$  (again, here, we emphasize the dependence on  $\omega \in \Omega$  for clarity, even though it was omitted above). First, note that for all  $u \in \mathbb{R}^d$ ,  $h(\cdot, u)$  is measurable, as the supremum of a countable family of measurable functions. Second, for all  $\omega \in \Omega$ , the function  $h(\omega, \cdot)$  is convex as the supremum of affine functions, and it only takes finite values: Indeed,  $C \subseteq B(t_0, \varepsilon)$  is bounded and  $f(\omega, \cdot)$  is continuous on  $B(t_0, \varepsilon)$ . Hence,  $h(\omega, \cdot)$  is continuous on  $\mathbb{R}^d$ . Therefore, since  $K$  is compact,  $\Gamma(\omega) \cap K \neq \emptyset$  if and only if  $\min_{u \in K} h(\omega, u) \leq 0$ , if and only if  $\inf_{u \in \tilde{K}} h(\omega, u) \leq 0$ , where  $\tilde{K}$  is a fixed, countable, dense subset of  $K$ . Therefore, we obtain  $\{\omega \in \Omega : \Gamma(\omega) \cap K \neq \emptyset\} = \{\omega \in \Omega : \inf_{u \in \tilde{K}} h(\omega, u) \leq 0\}$  which is measurable, since  $\inf_{u \in \tilde{K}} h(\cdot, u)$  is a measurable map.  $\square$

Finally, let us state an incredibly simple yet powerful result that shows that for convex functions, there is no need to apply any dominated convergence theorem in order to swap expectations and (sub-) gradients. It is very easy to check that if  $f_1$  and  $f_2$  are two convex functions on a convex set  $G_0 \subseteq \mathbb{R}^d$ , then for all  $t_0 \in G_0$ ,  $\partial f_1(t_0) + \partial f_2(t_0) \subseteq \partial(f_1 + f_2)(t_0)$ <sup>2</sup>. The following lemma shows that this fact still holds for generalized sums of convex functions.

**THEOREM 3.** *Let  $f$  be a random convex function defined on a convex set  $G_0 \subseteq \mathbb{R}^d$ . For all  $t \in \text{int}(G_0)$ , let  $g(t)$  be a measurable subgradient of  $f$  at  $t$ . Let  $p \geq 1$  be a real number and assume that for all  $t \in G_0$ ,  $f(t) \in L^p(\mathbb{P})$  and denote by  $F(t) = \mathbb{E}[f(t)]$ . Then,  $F$  is a convex function and for all  $t \in G_0$ ,  $g(t) \in L^p(\mathbb{P})$  and*

$$\mathbb{E}[g(t)] \in \partial F(t).$$

**PROOF.** Fix  $t_0 \in \text{int}(G_0)$  and let  $g(t_0)$  be a measurable subgradient of  $h$  at  $t_0$  (the existence of which is guaranteed by Theorem 3). In order to check that  $g(t_0) \in L^p(\mathbb{P})$ , it is necessary and sufficient to check that each of its  $d$  coordinates are in  $L^p(\mathbb{P})$  or, equivalently, that for all  $v \in \mathbb{R}^d$ ,  $|g(t_0)^\top v|^p$  is integrable. Fix an arbitrary  $v \in \mathbb{R}^d$  and let  $\varepsilon > 0$  be such that  $t_0 + \varepsilon v$  and  $t_0 - \varepsilon v$  are in  $G_0$  (such an  $\varepsilon$  exists because  $t_0 \in \text{int}(G_0)$ ). Then, by definition of subgradients,  $g(t_0)^\top v \leq \varepsilon^{-1}(f(t_0 + \varepsilon v) - f(t_0))$  and  $-g(t_0)^\top v \leq \varepsilon^{-1}(f(t_0 - \varepsilon v) - f(t_0))$ . That is,

$$|g(t_0)^\top v| \leq \max(\varepsilon^{-1}(f(t_0 + \varepsilon v) - f(t_0)), \varepsilon^{-1}(f(t_0 - \varepsilon v) - f(t_0))).$$

Since the right hand side is in  $L^p(\mathbb{P})$  by assumption, so is  $g(t_0)^\top v$ . The vector  $v$  was arbitrary, so we conclude that  $g(t_0) \in L^p(\mathbb{P})$ .

Now, for the rest of the proof, simply note that, again, by definition of subgradients,

$$f(t) \geq f(t_0) + g(t_0)^\top (t - t_0)$$

holds for all  $t \in G_0$ . Taking the expectation, which is linear, yields that

$$F(t) \geq F(t_0) + \mathbb{E}[g(t_0)]^\top (t - t_0)$$

which concludes the proof. □

**REMARK 2.**

- In fact, to obtain that  $g(t_0) \in L^p(\mathbb{P})$ , it would have been sufficient to assume that  $f(t) \in L^p(\mathbb{P})$  for all  $t \in B(t_0, \varepsilon)$ , for any arbitrary, small enough  $\varepsilon > 0$ .
- As a consequence of Theorem 3, if  $F$  is differentiable at  $t_0 \in \text{int}(G_0)$ , then  $\mathbb{E}[g(t_0)]$  does not depend on the choice of the measurable selection  $g(t_0)$  and it is automatically equal to  $\nabla F(t_0)$  (since  $\nabla F(t_0)$  is the only subgradient of  $F$  at  $t_0$ , in that case).
- In fact, Lemma 12 shows that if  $F$  is differentiable at some  $t_0 \in \text{int}(G_0)$ , then  $f$  is almost surely differentiable at  $t_0$ , so in that case, any measurable selection  $g(t_0)$  must satisfy  $g(t_0) = \nabla f(t_0)$  almost surely.
- To the best of our knowledge, the converse inclusion to Theorem 3 is unknown: Can all subgradients of  $F$  at  $t_0$  be written as  $\mathbb{E}[g(t_0)]$  for some measurable  $g(t_0) \in \partial f(t_0)$ ?

---

<sup>2</sup>The other inclusion is also true if  $G_0$  has non-empty interior but, perhaps surprisingly, requires a nontrivial argument.

### 3. CONSISTENCY

Consistency of empirical risk minimizers with a convex loss function is automatically granted in a strong sense, thanks to Lemma 1 which allows to localize the  $M$ -estimator, for large enough  $n$ , in an arbitrarily small neighborhood of the set of population minimizers with probability 1. In what follows, we consider a sequence  $(\hat{\theta}_n)_{n \geq 1}$  of random variables such that with probability 1, for all large enough  $n$ ,  $\hat{\theta}_n$  is a minimizer of  $\Phi_n$  on  $\Theta$ . Existence of such a sequence is granted by Theorem 1.

**THEOREM 4.** *Assume that  $\Theta^*$  is compact and non-empty. Then,  $d(\hat{\theta}_n, \Theta^*) \xrightarrow[n \rightarrow \infty]{} 0$  almost surely, as  $n \rightarrow \infty$ .*

The proof of this theorem can be found in [19] (the only difference here being that we do not assume that  $\Theta = \mathbb{R}^d$ ), and it is a direct consequence of Theorem 1 above.

**REMARK 3.** *Theorem 4 shows that any empirical minimizer becomes, with probability 1, arbitrarily close to the set of population minimizers  $\Theta^*$ . A converse statement is generally not true, that is, there can be elements of  $\Theta^*$  that may never be approached by any empirical minimizer. For instance, let  $E = \mathbb{R}^d$ ,  $\Theta = B(0, 1)$  and  $\phi(x, \theta) = x^\top \theta$ . Furthermore, assume that  $X_1$  has the standard normal distribution. Then,  $\Phi(\theta) = \mathbb{E}[X]^\top \theta = 0$  for all  $\theta \in \Theta$ , so  $\Theta^* = \Theta$ . However,  $\Phi_n(\theta) = \bar{X}_n^\top \theta$ , so with probability 1, the empirical minimizer is unique, given by  $\hat{\theta}_n = -\bar{X}_n / \|\bar{X}_n\|$ .*

### 4. ASYMPTOTIC DISTRIBUTION

In this section, we assume that  $\text{Argmin}_{\theta \in \Theta} \Phi(\theta)$  is a singleton and we denote by  $\theta^* = \text{argmin}_{\theta \in \Theta} \Phi(\theta)$ .

#### 4.1 Non-differentiable case

We first study asymptotic properties of  $\hat{\theta}_n$  without assuming differentiability of  $\Phi$  at  $\theta^*$ . That is,  $\partial\Phi(\theta^*)$  may not be a singleton.

The following useful property is fundamental in that case. Recall that for a non-empty convex subset  $C \subseteq \mathbb{R}^d$ , we denote by  $h_C : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$  its support function.

**PROPOSITION 1.** *Assume that  $\phi(\cdot, \theta) \in L^2(P)$  for all  $\theta \in \Theta_0$ . Let  $(\rho_n)_{n \geq 1}$  be any non-decreasing sequence of positive numbers diverging to  $\infty$  as  $n \rightarrow \infty$ . Then, for all  $\theta \in \Theta_0$  and  $t \in \mathbb{R}^d$ ,*

$$\rho_n(\Phi_n(\theta + t/\rho_n) - \Phi_n(\theta)) \xrightarrow[n \rightarrow \infty]{} h_{\partial\Phi(\theta)}(t)$$

*in probability.*

**PROOF.** Fix  $\theta \in \Theta_0$ . For all  $t \in \mathbb{R}^d$ , define

$$\begin{aligned} F_n(t) &= \rho_n \left( \Phi_n(\theta + t/\rho_n) - \Phi_n(\theta) - \frac{1}{n\rho_n} t^\top \sum_{i=1}^n g(X_i, \theta) \right) \\ &\quad - \rho_n \left( \Phi(\theta + t/\rho_n) - \Phi(\theta) - \frac{1}{\rho_n} t^\top \mathbb{E}[g(X_1, \theta)] \right). \end{aligned}$$

Write  $F_n(t) = \sum_{i=1}^n (Z_{i,n} - \mathbb{E}[Z_{i,n}])$  where  $Z_{i,n} = \frac{\rho_n}{n} (\phi(X_i, \theta + t/\rho_n) - \phi(X_i, \theta) - (1/\rho_n) t^\top g(X_i, \theta))$ , for all  $i = 1, \dots, n$ . Convexity of  $\phi(X_i, \cdot)$  yields that  $0 \leq Z_{i,n} \leq \frac{1}{n} t^\top (g(X_i, \theta + t/\rho_n) - g(X_i, \theta))$ , for all  $i = 1, \dots, n$ . By Theorem 3, each  $Z_{i,n}$ ,  $i = 1, \dots, n$ , is square-integrable. Hence, taking the square and the expectation in the last display,

$$\mathbb{E}[Z_{i,n}^2] \leq \frac{1}{n^2} \mathbb{E}[Y_n^2]$$

where  $Y_n = t^\top (g(X_1, \theta + t/\rho_n) - g(X_1, \theta))$ . Since  $(\rho_n)_{n \geq 1}$  is non-decreasing, Lemma 11 implies that the sequence  $(Y_n)_{n \geq 1}$  is non-increasing, yielding that  $\mathbb{E}[Z_{i,n}^2] \leq \frac{1}{n^2} \mathbb{E}[Y_1^2]$  and, by independence of  $X_1, X_2, \dots$ ,

$$\text{var} \left( \sum_{i=1}^n Z_{i,n} \right) = \sum_{i=1}^n \text{var}(Z_{i,n}) \leq \sum_{i=1}^n \mathbb{E}[Z_{i,n}^2] \leq \frac{\mathbb{E}[Y_1^2]}{n} \xrightarrow{n \rightarrow \infty} 0.$$

We conclude that  $F_n(t) \xrightarrow{n \rightarrow \infty} 0$  in  $L^2$  and, hence, in probability. Now, rewrite  $F_n(t)$  as

$$(3) \quad F_n(t) = \rho_n (\Phi_n(\theta + t/\rho_n) - \Phi_n(\theta)) - t^\top \left( \frac{1}{n} \sum_{i=1}^n g(X_i, \theta) - \mathbb{E}[g(X_1, \theta)] \right)$$

$$(4) \quad - \rho_n (\Phi(\theta + t/\rho_n) - \Phi(\theta)).$$

The law of large numbers yields that the term (3) converges to 0 in probability, and the term in (4) goes to  $d^+ \Phi(\theta; t)$  as  $n \rightarrow \infty$ . The result then follows from Lemma 9.  $\square$

As a consequence, we obtain the following theorem.

**THEOREM 5.** *Assume that  $\phi(\cdot, \theta) \in L^2(P)$  for all  $\theta \in \Theta_0$  and that  $0 \in \text{int}(\partial \Phi(\theta^*))$ . Then,  $\hat{\theta}_n = \theta^*$  with probability going to 1 as  $n \rightarrow \infty$ .*

Note that the assumption that  $0 \in \text{int}(\partial \Phi(\theta^*))$  readily implies that  $\theta^*$  must be the unique minimizer of  $\phi$  on  $\Theta$  and even on  $\Theta_0$ . It also implies that  $\Phi$  is not differentiable at  $\theta^*$ .

**PROOF.** Let  $(\rho_n)_{n \geq 1}$  be any non-decreasing sequence of positive numbers diverging to  $\infty$  as  $n \rightarrow \infty$ . Since  $\Theta_0$  is open, we can find  $r > 0$  such that  $B(\theta^*, r) \subseteq \Theta_0$ . For all  $n \geq 1$ , denote by  $T_n = \{t \in \mathbb{R}^d : \theta^* + t/\rho_n \in \Theta\} = \rho_n(\Theta - \theta^*)$ . Finally, set  $G_n(t) = \rho_n(\Phi_n(\theta^* + t/\rho_n) - \Phi_n(\theta^*))$ , for all  $t \in \mathbb{R}^d$  such that  $\theta^* + t/\rho_n \in \Theta_0$ . By definition of  $\hat{\theta}_n$ ,  $\hat{t}_n := \rho_n(\hat{\theta}_n - \theta^*)$  is a minimizer of  $G_n$  on  $T_n$  for all large enough  $n$ , with probability 1.

Now, fix  $\varepsilon > 0$ . Combining Proposition 1, Corollary 1 and Lemma 9, we get

$$\sup_{t \in B(0, \varepsilon)} |G_n(t) - h_{\partial \Phi(\theta^*)}(t)| \xrightarrow{n \rightarrow \infty} 0$$

in probability (note that  $B(0, \varepsilon) \subseteq \rho_n(\Theta_0 - \theta^*)$  for all large enough integers  $n$ ). Now, since  $0 \in \text{int}(\partial \Phi(\theta^*))$ , the quantity  $\eta := \min_{u \in \mathbb{R}^d : \|u\|=1} h_{\partial \Phi(\theta^*)}(u)$  is positive.

Assume that  $n$  is large enough so  $\sup_{t \in B(0, \varepsilon)} |G_n(t) - h_{\partial \Phi(\theta^*)}(t)| \leq \varepsilon \eta / 2$  with probability at least  $1 - \varepsilon$ . When this inequality is satisfied, we get that, for all  $t \in T_n$  with  $\|t\| = \varepsilon$ ,

$$\begin{aligned} G_n(t) &\geq h_{\partial \Phi(\theta^*)}(t) - \varepsilon \eta / 2 \\ &= \varepsilon h_{\partial \Phi(\theta^*)}(t/\varepsilon) - \varepsilon \eta / 2 \quad \text{by positive homogeneity of } h_{\partial \Phi(\theta^*)} \\ &\geq \varepsilon \eta - \varepsilon \eta / 2 \quad \text{by definition of } \eta \\ &> \varepsilon \eta / 2 \\ &> 0 = G_n(0) \end{aligned}$$

yielding, thanks to Lemma 1, that  $\|\hat{t}_n\|$  cannot be larger than  $\varepsilon$ . Hence, we have shown that for all large enough  $n$ , it holds with probability at least  $1 - \varepsilon$  that  $\|\rho_n(\hat{\theta}_n - \theta^*)\| \leq \varepsilon$ . That is,  $\rho_n(\hat{\theta}_n - \theta^*) \xrightarrow{n \rightarrow \infty} 0$  in probability. Since this must hold for any positive, non-decreasing sequence  $(\rho_n)_{n \geq 1}$  diverging to  $\infty$  as  $n \rightarrow \infty$ , Lemma 25 implies the desired statement.  $\square$

Let  $C$  be the support cone to  $\Theta$  at  $\theta^*$ . Recall that the first order condition (Lemma 10) yields that  $C \subseteq h_{\partial\Phi(\theta^*)}^{-1}([0, \infty))$ . The next result extends Theorem 5.

**THEOREM 6.** *Assume that  $\phi(\cdot, \theta) \in L^2(P)$  for all  $\theta \in \Theta_0$  and that  $h_{\partial\Phi(\theta^*)}(t) > 0$  for all  $t \in C \setminus \{0\}$ . Then, with probability going to 1 as  $n \rightarrow \infty$ ,  $\hat{\theta}_n = \theta^*$ .*

The assumption of the theorem is that the two closed, convex cones  $C$  and  $\{t \in \mathbb{R}^d : h_{\partial\Phi(\theta^*)}(t) \leq 0\}$  have a trivial intersection. Note that, by the first order condition at  $\theta^*$ , this intersection must always be included in the boundary of  $C$ . In other words, the assumption of the theorem is that all (non-zero) vectors in  $C$  are directions of strict, linear increase of the population risk  $\Phi$ .

**PROOF.** A consequence of the assumption of the theorem is that for all  $\varepsilon > 0$ ,  $\{t \in C : h_{\partial\Phi(\theta^*)}(t) \leq \varepsilon\}$  is compact. Indeed, it is closed, since  $C$  is closed and  $h_{\partial\Phi(\theta^*)}$  is continuous. Moreover, the set  $\{t \in C : \|t\| = 1\}$  is compact, so by continuity of  $h_{\partial\Phi(\theta^*)}$ , there is some  $t_0 \in C$  with  $\|t_0\| = 1$  satisfying, for all  $t \in C \setminus \{0\}$ ,  $h_{\partial\Phi(\theta^*)}(t) \geq \|t\| h_{\partial\Phi(\theta^*)}(t_0)$ . The assumption of the theorem implies that  $h_{\partial\Phi(\theta^*)}(t_0) > 0$ . Finally,  $\{t \in C : h_{\partial\Phi(\theta^*)}(t) \leq \varepsilon\}$  is bounded, since it is included in  $B(0, \varepsilon/h_{\partial\Phi(\theta^*)}(t_0))$ .

Now, let  $(\rho_n)_{n \geq 1}$  be an arbitrary non-decreasing sequence of positive numbers, diverging to  $\infty$  as  $n \rightarrow \infty$  and fix  $\varepsilon > 0$ . Proposition 1, Corollary 1 and Lemma 9, yield that  $\sup_{t \in C : h_{\partial\Phi(\theta^*)}(t) \leq \varepsilon} |G_n(t) - h_{\partial\Phi(\theta^*)}(t)| \xrightarrow[n \rightarrow \infty]{} 0$  in probability, where we set  $G_n(t) = \rho_n(\Phi_n(\theta^* + t/\rho_n) - \Phi_n(\theta^*))$  as in the proof of Theorem 5. Let  $n$  be large enough so  $\sup_{t \in C : h_{\partial\Phi(\theta^*)}(t) \leq \varepsilon} |G_n(t) - h_{\partial\Phi(\theta^*)}(t)| \leq \varepsilon/2$  with probability at least  $1 - \varepsilon$ . Then, with probability at least  $1 - \varepsilon$ , it holds simultaneously for all  $t \in T_n = \rho_n(\Theta - \theta^*)$  with  $h_{\partial\Phi(\theta^*)}(t) = \varepsilon$ , that

$$G_n(t) \geq h_{\partial\Phi(\theta^*)}(t) - \varepsilon/2 = \varepsilon/2 > 0 = G_n(0)$$

so, by Lemma 1, any minimizer  $\hat{t}_n$  of  $G_n$  on  $T_n$  satisfies  $h_{\partial\Phi(\theta^*)}(\hat{t}_n) \leq \varepsilon$ . In particular, we obtain, for all large enough  $n$ , that with probability at least  $1 - \varepsilon$ ,

$$0 \leq h_{\partial\Phi(\theta^*)}(\rho_n(\hat{\theta}_n - \theta^*)) = \rho_n h_{\partial\Phi(\theta^*)}(\hat{\theta}_n - \theta^*) \leq \varepsilon$$

where the first inequality follows from the first order condition for  $\Phi$  at  $\theta^*$  (Lemma 10). That is  $\rho_n h_{\partial\Phi(\theta^*)}(\hat{\theta}_n - \theta^*) \xrightarrow[n \rightarrow \infty]{} 0$ . Since the sequence  $(\rho_n)_{n \geq 1}$  was arbitrary, Lemma 25 yields that  $h_{\partial\Phi(\theta^*)}(\hat{\theta}_n - \theta^*) = 0$  with probability going to 1 as  $n \rightarrow \infty$ . Since  $\hat{\theta}_n - \theta^* \in C$ , this means that  $\hat{\theta}_n - \theta^* = 0$  with probability going to 1 as  $n \rightarrow \infty$ , which is the desired statement.  $\square$

**REMARK 4.** *Results of this section rely on Proposition 1, which imposes square-integrability of the loss function. We do not know whether the same results could be proved under weaker assumptions.*

Now, to obtain a more precise asymptotic description of  $\hat{\theta}_n$  when  $\Phi$  is differentiable at  $\theta^*$  (this could be the case in Theorem 6, with  $\nabla\Phi(\theta^*)^\top t > 0$  for all  $t \in C \setminus \{0\}$ , but not in Theorem 5), we will assume the existence of second order derivatives for  $\Phi$  at  $\theta^*$ . This is the object of the next section.

## 4.2 Differentiable case

Let us first state the main result of this section.

**THEOREM 7.** *Let  $g : E \times \Theta_0 \rightarrow \mathbb{R}^d$  be a measurable selection of subgradients of  $\phi$ . Assume the following:*

- (i)  $\Phi$  is twice differentiable at  $\theta^*$  and  $S := \nabla^2 \Phi(\theta^*)$  is positive definite;
- (ii)  $g(\cdot, \theta^*) \in L^2(P)$ ;
- (iii)  $\pi_{\Theta-\theta^*}^S$  has directional derivatives at  $-S^{-1} \nabla \Phi(\theta^*)$ .

Then,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow[n \rightarrow \infty]{} d^+ \pi_{\Theta-\theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); Z)$$

in distribution, where  $Z \sim \mathcal{N}_d(0, S^{-1} B S^{-1})$  and  $B = \text{var}(g(X_1, \theta^*))$ .

REMARK 5 (on the assumptions of the theorem).

- (i) Second differentiability of  $\Phi$  at  $\theta^*$  is not a strong restriction, since all convex functions are twice differentiable almost everywhere in the interior of their domains [1]. The assumption that  $\nabla^2 \Phi(\theta^*)$  is definite positive is made in order to obtain  $n^{-1/2}$  convergence rate. This assumption could be relaxed, yielding slower rates under further, technical assumptions on higher order derivatives on  $\Phi$ . In this work, we choose to focus on the  $n^{-1/2}$  rate because it only requires minimal, easy to check, non-restrictive smoothness assumptions.
- (ii) Existence of the map  $g$  is guaranteed by Theorem 3. Moreover, the first assumption on  $\Phi$  implies that it is differentiable at  $\theta^*$ , so by Lemma 12,  $\phi(X_1, \cdot)$  is almost surely differentiable at  $\theta^*$  yielding that  $g(x, \theta^*) = \nabla(\phi(x, \cdot))(\theta^*)$  for  $P$ -almost all  $x \in E$ . Theorem 3 also ensures that it is sufficient that  $\phi(\cdot, \theta) \in L^2(P)$  for all  $\theta \in \Theta_0$  for the second assumption to hold. In fact, a straightforward adaptation of Theorem 3 shows that it is even enough to only assume that  $\phi(\cdot, \theta) \in L^2(P)$  for all  $\theta$  in any arbitrarily small neighborhood of  $\theta^*$ . Note that this does not require a uniform domination of  $\phi$  or its derivatives/subgradients in any neighborhood of  $\theta^*$  but, rather, a pointwise integrability condition of order 0 (that is, on  $\phi$  itself).
- (iii-a) Directional differentiability of  $\pi_{\Theta-\theta^*}^S$  is not a strong restriction in the sense that,  $\pi_{\Theta-\theta^*}^S$  being non-expansive (see Lemma 13) it is automatically differentiable almost everywhere by Rademacher's theorem [16, Section 3.1.6, p. 216]. In the appendix (Section C), we present several sufficient conditions that guarantee the existence of directional derivatives of  $\pi_K^S$  for a convex set  $K$ , at a direction  $u$ , which, in practice, are easily checked (e.g.,  $u \in K$ , or  $u \notin K$  and  $\partial K$  is smooth at  $\pi_K(u)$ , or  $K$  is defined by finitely many linear convex constraints, etc.). By an obvious linear change of variables, it is clear that the existence of a directional derivative of  $\pi_{\Theta-\theta^*}^S$  at  $-S^{-1} \nabla \Phi(\theta^*)$  in a direction  $z \in \mathbb{R}^d$  is equivalent to the existence of a directional derivative of  $\pi_{S^{1/2}(\Theta-\theta^*)}$  at  $-S^{-1/2} \nabla \Phi(\theta^*)$  in the direction  $S^{1/2} z$ . Then, simple algebra yields that

$$d^+ \pi_{\Theta-\theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); z) = S^{-1/2} d^+ \pi_{S^{1/2}(\Theta-\theta^*)}(-S^{-1/2} \nabla \Phi(\theta^*); S^{1/2} z).$$

Recall that  $(\theta - \theta^*)^\top \nabla \Phi(\theta^*) \geq 0$  for all  $\theta \in \Theta$ : This is granted by the first order condition at  $\theta^*$  (Lemma 10). That is,  $-\nabla \Phi(\theta^*)$  is in the normal cone to  $\Theta$  at  $\theta^*$  or, equivalently,  $-S^{-1/2} \nabla \Phi(\theta^*)$  is in the normal cone to  $S^{1/2}(\Theta - \theta^*)$  at 0.

REMARK 6 (on the conclusion of the theorem).

- Lemma 20 yields that for any  $z \in \mathbb{R}^d$ ,  $d^+ \pi_{\Theta-\theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); z) \in C_{S^{-1} \nabla \Phi(\theta^*)}^S = C_{\nabla \Phi(\theta^*)}$  where  $C$  is the support cone to  $\Theta$  at  $\theta^*$ . When  $\nabla \Phi(\theta^*)^\top t > 0$  for all  $t \in C \setminus \{0\}$  (that is,  $-\nabla \Phi(\theta^*)$  is in the interior of the normal cone to  $\Theta$  at  $\theta^*$ ),  $C_{\nabla \Phi(\theta^*)} = \{0\}$ ,  $d^+ \pi_{\Theta-\theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); \cdot) = 0$  so Theorem 7 yields that  $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow[n \rightarrow \infty]{} 0$  in distribution: This was already a (rather weak) consequence of Theorem 6.
- If  $\theta^* \in \text{int}(\Theta)$ , then the first order condition (Lemma 10) yields that  $\nabla \Phi(\theta^*) = 0$  and,  $d^+ \pi_{\Theta-\theta^*}^S(0; \cdot)$  is simply the identity map. Therefore, Theorem 7 says that  $\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow[n \rightarrow \infty]{} Z$



in distribution. In that case, Theorem 4 implies that, with probability 1, for all large enough  $n$ ,  $\hat{\theta}_n \in \text{int}(\Theta)$ . Hence, with probability 1, for all large enough  $n$ ,  $\hat{\theta}_n$  (the constrained  $M$ -estimator) is also a solution to the unconstrained optimization problem  $\min_{\theta \in \Theta_0} \Phi_n(\theta)$ , and we recover Haberman's theorem [19, Theorem 6.1].

- In fact, Theorem 7 also encompasses the unconstrained case, by taking  $\Theta = \Theta_0 = \mathbb{R}^d$ . If  $\Theta_0$  is a strict open subset of  $\mathbb{R}^d$ , one can also consider an unconstrained  $M$ -estimator  $\tilde{\theta}_n$  on the open set  $\Theta_0$ , that is, a minimizer of  $\Phi_n$  on  $\Theta_0$ . Assume that  $\theta^*$  is the unique minimizer of  $\Phi$  on the open set  $\Theta_0$  and let  $\Theta$  be any closed subset of  $\Theta_0$  containing  $\theta^*$  in its interior (e.g., take  $\Theta = B(\theta^*, \varepsilon)$  for any small enough  $\varepsilon$ ). Then, a straight adaptation of Theorem 4 yields that  $\tilde{\theta}_n \xrightarrow[n \rightarrow \infty]{} \theta^*$  almost surely, so  $\tilde{\theta}_n \in \Theta$  for all large enough  $n$ , with probability 1. That is,  $\tilde{\theta}_n$  eventually coincides with a constrained  $M$ -estimator and, hence, also satisfies the conclusion of Theorem 7, with  $d^+ \pi_{\Theta-\theta^*}^S(0; \cdot)$  being the identity map (note that in the case  $\Theta = \Theta_0 = \mathbb{R}^d$ , we necessarily have that  $\nabla \Phi(\theta^*) = 0$ ).
- If the boundary of  $\Theta$  is  $C^2$  in a neighborhood of  $\theta^*$  (that is, it can be locally represented as the graph of a  $C^2$  mapping from  $\mathbb{R}^{d-1}$  to  $\mathbb{R}$ ) and  $\nabla \Phi(\theta^*) \neq 0$ , then, Lemma 15 yields that  $\sqrt{n}(\hat{\theta}_n - \theta^*)$  converges in distribution to a Gaussian distribution that is supported in the linear hyperplane that is parallel to the (unique) supporting hyperplane to  $\Theta$  at  $\theta^*$ .
- Lemmas 23 and 24 imply that for all  $t, t' \geq 0$  with  $t' > t$ ,

$$(5) \quad \|d^+ \pi_{\Theta-\theta^*}^S(-t' S^{-1} \nabla \Phi(\theta^*); Z)\|_S \leq \|d^+ \pi_{\Theta-\theta^*}^S(-t S^{-1} \nabla \Phi(\theta^*); Z)\|_S$$

almost surely. This can be interpreted as follows. First, note that the set  $\Theta$  can represent some constraints that are imposed by a specific application, or it can represent a model (e.g., if it is believed that the global minimizer of  $\Phi$  lies in  $\Theta$ ). In the latter case, the model is misspecified if the global minimizer of  $\Phi$  is not in  $\Theta$ , that is, if  $\nabla \Phi(\theta^*) \neq 0$ . In other words, the vector  $\nabla \Phi(\theta^*)$  (or its rescaled version  $S^{-1} \nabla \Phi(\theta^*)$ ) can be used to quantify the amount of model misspecification. In that regard, (5) suggests that more misspecification yields better asymptotic error (we do not account for any misspecification bias here). In (5),  $t = 0$  can be thought of as corresponding to the well-specified case. This will be illustrated in the examples below.

- As a consequence of Theorem 7, the mean squared error of  $\hat{\theta}_n$  satisfies

$$(6) \quad \liminf_{n \rightarrow \infty} n \mathbb{E}[\|\hat{\theta}_n - \theta^*\|_S^2] \geq \mathbb{E}[\|d^+ \pi_{\Theta-\theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); Z)\|_S^2]$$

(we do not know, in general, whether this is in fact an equality, with the  $\liminf$  being a simple limit, see the open question below). The right hand side can be interpreted as a local measure of the statistical complexity of  $\Theta$  around  $\theta^*$ , relative to the (population) loss function  $\Phi$ . The statistical dimension (or Gaussian width) of a non-empty, closed, convex set  $G \subseteq \mathbb{R}^d$  is measured as  $\mathbb{E}[\|\pi_G(Z)\|^2]$  where  $Z \sim \mathcal{N}_d(0, I_d)$ , see [3] (in our case, we need to account for a scaling given by  $S^{-1}$  and  $B$  in the covariance matrix of  $Z$ ). In (6), we do not have a projection, but the directional derivative of a projection. The right hand side of (6) can rather be seen as a statistical dimension at an infinitesimal scale. We can refer, for instance, to [11] who studied least squares under convex constraint, and proved that the statistical dimension at a fixed scale drives the statistical error. A similar phenomenon has also been studied for constrained  $M$ -estimators in a more general setup [35]. Recall, however, that except in specific cases (see Section C in the appendix),  $d^+ \pi_{\Theta-\theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); \cdot)$  is not the projection onto a convex set.

- It is worth mentioning some further important properties of  $\Pi := d^+ \pi_{\Theta-\theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); \cdot)$ . As we have noted above, in general, it is not the projection onto a convex cone. Nevertheless,

it shares similar properties as the projection onto a convex cone. Indeed, by Lemma 21, it satisfies the following properties:

- $\Pi(\lambda z) = \lambda \Pi(z)$ , for all  $\lambda \geq 0$  and  $z \in \mathbb{R}^d$  (positive homogeneity);
- $\|\Pi(z') - \Pi(z)\|_S \leq \|z' - z\|_S^2$  (non-expansiveness);
- $\langle \Pi(z') - \Pi(z), z' - z \rangle_S \geq \|\Pi(z') - \Pi(z)\|_S^2 \geq 0$  for all  $z, z' \in \mathbb{R}^d$  (firm monotonicity).

Note that non-expansiveness is implied by firm monotonicity. Such maps satisfying the last two properties above have been studied extensively [57]. Moreover, [43, Proposition 2.1] implies that  $\Pi$  is the gradient of a convex function.

Now, let us look at some applications of Theorem 7.

**EXAMPLE 1** (Constrained mean estimation). Let  $X_1, X_2, \dots$  be iid random vectors with two moments<sup>3</sup> and  $\Theta \subseteq \mathbb{R}^d$  be a non-empty, closed, convex set. Consider the loss function  $\phi(x, \theta) = (1/2)\|x - \theta\|^2$ ,  $x, \theta \in \mathbb{R}^d$ . Then,  $\theta^* = \pi_\Theta(\mathbb{E}[X_1])$  is the unique minimizer of  $\Phi$  on  $\Theta$  and  $\hat{\theta}_n = \pi_\Theta(\bar{X}_n)$  where  $\bar{X}_n = n^{-1}(X_1 + \dots + X_n)$ , for all  $n \geq 1$ . Consistency, which is a consequence of Theorem 4, also follows directly from the strong law of large numbers, together with continuity of  $\pi_\Theta$  (since it is non-expansive). For asymptotic normality, we obtain, from Theorem 7, that

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow[n \rightarrow \infty]{} d^+ \pi_{\Theta - \theta^*}(\mathbb{E}[X_1] - \theta^*; Z) = d^+ \pi_\Theta(\mathbb{E}[X_1]; Z)$$

in distribution, where  $Z \sim \mathcal{N}_d(0, \text{var}(X_1))$  (in this example,  $S = I_d$ ). In this simple case, this result can also be obtained using the central limit theorem, combined with the delta method<sup>4</sup>.

Here, it is clear that misspecification is favorable for the asymptotic error: For instance, if  $\Theta - \theta^*$  is a convex cone and  $\mathbb{E}[X_1] - \theta^*$  is in the interior of the normal cone to  $\Theta$  at  $\theta^*$  (in particular,  $\theta^* \neq \mathbb{E}[X_1]$ ), then, Theorem 5 yields that  $\hat{\theta}_n = \theta^*$  with probability going to 1 as  $n \rightarrow \infty$ .

**EXAMPLE 2** (Constrained least squares). Let  $(X_1, Y_1), (X_2, Y_2), \dots$  be iid random pairs in  $\mathbb{R}^d \times \mathbb{R}$ . Assume that  $X_1$  has four moments,  $\mathbb{E}[X_1] = 0$ ,  $S := \mathbb{E}[X_1 X_1^\top]$  is definite positive,  $Y_1 - X_1^\top \theta_0$  is independent of  $X_1$  and has the centered Gaussian distribution with variance  $\sigma^2 > 0$  for some  $\theta_0 \in \mathbb{R}^d$  and  $\sigma^2 > 0$ . Let  $\phi(x, y, \theta) = 1/2(y - x^\top \theta)^2$ , for all  $x \in \mathbb{R}^d, y \in \mathbb{R}$  and  $\theta \in \mathbb{R}^d$ . Then, for all  $\theta \in \mathbb{R}^d$ ,

$$\Phi(\theta) = \frac{1}{2} \|\theta - \theta_0\|_S^2 + \sigma^2.$$

Let  $\Theta \subseteq \mathbb{R}^d$  be a non-empty, closed, convex subset of  $\mathbb{R}^d$  (here,  $\Theta_0 = \mathbb{R}^d$ ). Then,  $\text{Argmin}_{\theta \in \Theta} \Phi(\theta) = \{\pi_\Theta^S(\theta_0)\}$  and, provided that  $\pi_\Theta$  has directional derivatives at  $\theta_0$ , the least square estimator  $\hat{\theta}_n$ , defined as any minimizer on  $\Theta$  of  $\Phi_n(\theta) = n^{-1} \sum_{i=1}^n (Y_i - X_i^\top \theta)^2$ ,  $\theta \in \mathbb{R}^d$ , satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow[n \rightarrow \infty]{} d^+ \pi_{\Theta - \theta^*}^S(\theta_0 - \theta^*; Z) = d^+ \pi_\Theta^S(\theta_0; Z)$$

in distribution, where  $Z \sim \mathcal{N}_d(0, S^{-1} B S^{-1})$  and

$$\begin{aligned} B &= \text{var}((Y_1 - X_1^\top \theta^*) X_1) \\ &= \text{var}((Y_1 - X_1^\top \theta_0) X_1 + X_1^\top (\theta^* - \theta_0) X_1) \\ &= \mathbb{E}[(X_1^\top (\theta_0 - \theta^*))^2 X_1 X_1^\top] + \sigma^2 S. \end{aligned}$$

<sup>3</sup>In fact, one moment is enough if one rather uses the loss function  $\phi(x, \theta) = \|x - \theta\|^2 - \|x\|^2$ ,  $x, \theta \in \mathbb{R}^d$

<sup>4</sup>Delta method requires Hadamard directional differentiability of  $\pi_{\Theta - \theta^*}$  at  $\mathbb{E}[X_1] - \theta^*$ . This is readily implied by the existence of directional derivatives together with non-expansiveness of  $\pi_{\Theta - \theta^*}$

EXAMPLE 3 (Geometric median). *Let  $X_1, X_2, \dots$  be iid random vectors with one moment<sup>5</sup>. Consider the loss function  $\phi(x, \theta) = \|x - \theta\|$ ,  $x, \theta \in \mathbb{R}^d$ . Then,  $\theta^*$  is any geometric median and  $\hat{\theta}_n$  is any empirical geometric median. Here, in the unconstrained case, we recover standard results for geometric median M-estimation, provided that the distribution of  $X_1$  is not supported on an affine line (this guarantees uniqueness of  $\theta^*$ ) and that  $1/\|X_1 - \theta^*\|$  is integrable (this guarantees that  $\Phi$  is twice differentiable at  $\theta^*$  with positive definite Hessian), see, e.g., [28].*

PROOF OF THEOREM 7. Recall that we denote by  $S = \nabla^2 \Phi(\theta^*)$ , which is a symmetric, positive definite matrix, by assumption.

First, since  $\Theta_0$  is open, there exists some  $r > 0$  such that  $B_S(\theta^*, r) \subseteq \Theta_0$ . Fix some  $R > 0$ , whose value will be determined later, and let  $n \geq 1$  be any integer that is large enough so  $R/\sqrt{n} \leq r$ . For all such integers  $n$ , let  $F_n$  be the random function defined on  $B(0, R)$  by

$$F_n(t) = n\left(\Phi_n(\theta^* + t/\sqrt{n}) - \Phi_n(\theta^*)\right) - \left(\frac{t^\top}{\sqrt{n}} \sum_{i=1}^n g(X_i, \theta^*) + \frac{1}{2} t^\top \nabla^2 \Phi(\theta^*) t\right)$$

for all  $t \in B_S(0, R)$ . This is a random convex function. Our first goal is to prove that  $F_n$  converges pointwise (and hence, by Corollary 1, uniformly on the compact set  $B_S(0, R)$ ) to zero in probability. From this, we will then obtain that any minimizer of the first term (one of which is given by  $\sqrt{n}(\hat{\theta}_n - \theta^*)$  for large enough  $n$ , with probability 1) is close to the unique minimizer of the second, quadratic term.

Fix  $t \in B_S(0, R)$  and  $n \geq 1$ . For  $i = 1, \dots, n$ , let  $Z_{i,n} = \phi(X_i, \theta^* + n^{-1/2}t) - \phi(X_i, \theta^*) - n^{-1/2}t^\top g(X_i, \theta^*)$ . By definition of subgradients,

$$0 \leq Z_{i,n} \leq n^{-1/2}t^\top (g(X_i, \theta^* + n^{-1/2}t) - g(X_i, \theta^*)).$$

Squaring and taking the expectation yields that

$$(7) \quad \mathbb{E}[Z_{i,n}^2] \leq n^{-1} \mathbb{E}\left[\left(t^\top (g(X_1, \theta^* + n^{-1/2}t) - g(X_1, \theta^*))\right)^2\right]$$

(we replaced  $i$  with 1 in the right hand side because the  $X_i$ 's are iid). Let  $Y_n := t^\top (g(X_1, \theta^* + n^{-1/2}t) - g(X_1, \theta^*))$ . As mentioned above,  $Y_n \geq 0$ . Moreover, for  $n \geq 1$ , letting  $u = \theta^* + t/\sqrt{n}$  and  $v = \theta^* + t/\sqrt{n+1}$ ,

$$\begin{aligned} Y_n - Y_{n+1} &= t^\top (g(X_1, u) - g(X_1, v)) \\ &= (1/\sqrt{n} - 1/\sqrt{n+1})^{-1} (u - v)^\top (g(X_1, u) - g(X_1, v)) \\ &\geq 0 \end{aligned}$$

by Lemma 11. So the sequence  $(Y_n)_{n \geq 1}$  is non-increasing. Hence,  $Y_n$  converges almost surely to some non-negative random variable  $Y$ . By monotone convergence (noting that  $Y_1$  is integrable), this implies that

$$(8) \quad \mathbb{E}[Y_n] \xrightarrow{n \rightarrow \infty} \mathbb{E}[Y].$$

However, for all  $n \geq 1$ ,  $\mathbb{E}[Y_n] = t^\top (w_n - \nabla \Phi(\theta^*))$  where  $w_n \in \partial \Phi(\theta^* + t/\sqrt{n})$ , by Lemma 6. Lemma 7 yielding that  $w_n \xrightarrow{n \rightarrow \infty} w$ , we obtain that  $\mathbb{E}[Y_n] \xrightarrow{n \rightarrow \infty} 0$ . Together with (8), this shows that  $\mathbb{E}[Y] = 0$

<sup>5</sup>Similarly to the first example, one need not assume the existence of one moment if the loss function is replaced with  $\phi(x, \theta) = \|x - \theta\| - \|x\|$ ,  $x, \theta \in \mathbb{R}^d$ .

and, hence, because  $Y \geq 0$ , that  $Y = 0$  almost surely. Therefore, again by monotone convergence (noting, this time, that  $Y_1^2$  is integrable),  $\mathbb{E}[Y_n^2] \xrightarrow{n \rightarrow \infty} \mathbb{E}[Y^2] = 0$ .

Combined with (7) and using independence of  $Z_{1,n}, \dots, Z_{n,n}$ , we obtain that

$$(9) \quad \text{var} \left( \sum_{i=1}^n Z_{i,n} \right) = \sum_{i=1}^n \text{var}(Z_{i,n}) \leq \sum_{i=1}^n \mathbb{E}[Z_{i,n}^2] \leq \mathbb{E}[Y_n^2] \xrightarrow{n \rightarrow \infty} 0.$$

Therefore, by Chebychev's inequality,  $\sum_{i=1}^n (Z_{i,n} - \mathbb{E}[Z_{i,n}]) \xrightarrow{n \rightarrow \infty} 0$  in probability, that is,

$$n \left( \Phi_n(\theta^* + n^{-1/2}t) - \Phi_n(\theta^*) \right) - n^{-1/2}t^\top \sum_{i=1}^n g(X_i, \theta^*) - n \left( \Phi(\theta^* + n^{-1/2}t) - \Phi(\theta^*) - n^{-1/2}t^\top \nabla \Phi(\theta^*) \right) \xrightarrow{n \rightarrow \infty} 0$$

in probability. Now, since we have assumed that  $\Phi$  is twice differentiable at  $\theta^*$ , we finally obtain that

$$(10) \quad F_n(t) \xrightarrow{n \rightarrow \infty} 0$$

in probability, for all  $t \in B_S(0, R)$ , as desired.

For all integers  $n \geq 1$ , let  $T_n = \{t \in \mathbb{R}^d : \theta^* + n^{-1/2}t \in \Theta\} = n^{1/2}(\Theta - \theta^*) \subseteq T$  and  $S_n = \{t \in \mathbb{R}^d : \theta^* + n^{-1/2}t \in \Theta_0\} = n^{1/2}(\Theta_0 - \theta^*)$ . Then,  $T_n$  is a closed subset of  $S_n$ . Moreover, since  $\theta^* \in \Theta_0$  and  $\Theta_0$  is open,  $B_S(0, R) \subseteq S_n$  for all large enough integers  $n$  (recall that  $R > 0$  is some fixed number, whose value is still to be determined). Define the maps

$$\hat{G}_n : t \in S_n \mapsto n \left( \Phi_n(\theta^* + n^{-1/2}t) - \Phi_n(\theta^*) \right)$$

and

$$G_n : t \in \mathbb{R}^d \mapsto n^{-1/2}t^\top \sum_{i=1}^n g(X_i, \theta^*) + \frac{1}{2}t^\top \nabla^2 \Phi(\theta^*)t.$$

As per these definitions,  $F_n = \hat{G}_n - G_n$ , so, (10) and Corollary 1 yield that

$$(11) \quad \sup_{t \in B_S(0, R)} |\hat{G}_n(t) - G_n(t)| \xrightarrow{n \rightarrow \infty} 0$$

in probability.

Moreover,  $\hat{t}_n := n^{1/2}(\hat{\theta}_n - \theta^*)$  is a minimizer of  $\hat{G}_n$  on  $T_n$ , by definition of the empirical risk minimizer  $\hat{\theta}_n$ .

Now, denote by  $Z_n = n^{-1/2}S^{-1} \sum_{i=1}^n g(X_i, \theta^*) - \nabla \Phi(\theta^*)$  and for all  $t \in \mathbb{R}^d$ , rewrite  $G_n(t)$  as

$$\begin{aligned} G_n(t) &= n^{-1/2}t^\top \sum_{i=1}^n g(X_i, \theta^*) + \frac{1}{2}t^\top \nabla^2 \Phi(\theta^*)t \\ &= \langle n^{-1/2}S^{-1} \sum_{i=1}^n g(X_i, \theta^*), t \rangle_S + \frac{1}{2}\|t\|_S^2 \\ &= \langle Z_n + \sqrt{n}S^{-1}\nabla \Phi(\theta^*), t \rangle_S + \frac{1}{2}\|t\|_S^2 \\ &= \frac{1}{2}\|t + Z_n + \sqrt{n}S^{-1}\nabla \Phi(\theta^*)\|_S^2 - \|Z_n + \sqrt{n}S^{-1}\nabla \Phi(\theta^*)\|_S^2. \end{aligned}$$

It is now clear that  $G_n$  has a unique minimizer on  $T_n$ , which we denote by  $t_n^*$  and which is given by

$$t_n^* = \pi_{T_n}^S(-Z_n - \sqrt{n}S^{-1}\nabla \Phi(\theta^*)).$$

Now, our goal is twofold. First, to study the asymptotic behavior of  $t_n^*$  and show that it converges in distribution, as  $n \rightarrow \infty$ . Second, to check, based on (11), that  $\hat{t}_n$  approaches  $t_n^*$  as  $n \rightarrow \infty$ , that is,  $\hat{t}_n - t_n^*$  converges in probability to 0. Using Slutsky's theorem, these two facts will imply convergence in distribution of  $\hat{t}_n$ .

*Asymptotic behavior of  $t_n^*$ .*

First, by the central limit theorem, we have that  $Z_n \xrightarrow[n \rightarrow \infty]{} Z$  in distribution, where  $Z$  is a centered Gaussian random variable with covariance matrix given by  $S^{-1} \text{var}(g(X_1, \theta^*)) S^{-1}$ .

By Skorohod representation theorem (see [25, Theorem 5.31] for instance), one may assume that  $Z_n$  converges almost surely to  $Z$ . Since  $\pi_C^S$  is non-expansive by Lemma 13, it holds that  $t_n^* - \pi_{T_n}^S(-Z - \sqrt{n} S^{-1} \nabla \Phi(\theta^*))$  converges to 0 almost surely. Moreover,

$$\begin{aligned} \pi_{T_n}^S(-Z - \sqrt{n} S^{-1} \nabla \Phi(\theta^*)) &= \pi_{\sqrt{n}(\Theta - \theta^*)}^S(-Z - \sqrt{n} S^{-1} \nabla \Phi(\theta^*)) \\ &= \sqrt{n} \pi_{\Theta - \theta^*}^S(-n^{-1/2} Z - S^{-1} \nabla \Phi(\theta^*)) \\ &\xrightarrow[n \rightarrow \infty]{} d^+ \pi_{\Theta - \theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); -Z) \end{aligned}$$

almost surely, using the third assumption of the theorem. Therefore, we conclude that  $t_n^* \xrightarrow[n \rightarrow \infty]{} d^+ \pi_{\Theta - \theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); -Z)$  almost surely and, hence, in distribution. The desired results follows, since  $Z$  and  $-Z$  are identically distributed.

*Convergence in probability of  $\hat{t}_n - t_n^*$  to 0.*

Fix  $\varepsilon > 0$ . Since the sequence  $(t_n^*)_{n \geq 1}$  converges in distribution (see the previous paragraph), it is tight, that is, there must exist some  $M > 0$  such that for all  $n \geq 1$ ,  $P(\|t_n^*\|_S \leq M) \geq 1 - \varepsilon$ . Let  $K = B_S(0, M + \varepsilon)$  and fix some  $\eta > 0$  to be chosen below. (11) yields that for all large enough  $n \geq 1$ ,  $\sup_{t \in K} |\hat{G}_n(t) - G_n(t)| \leq \eta$  with probability at least  $1 - \varepsilon$ . Therefore, by the union bound, for all large enough  $n \geq 1$ , it holds with probability at least  $1 - 2\varepsilon$  that simultaneously for all  $t \in T_n$  with  $\|t - t_n^*\|_S = \varepsilon$ ,

$$\begin{aligned} \hat{G}_n(t) &\geq G_n(t) - \eta \\ &\geq G_n(t_n^*) + \frac{\varepsilon^2}{2} - \eta \\ &\geq \hat{G}_n(t_n^*) - \eta + \frac{\varepsilon^2}{2} - \eta. \end{aligned}$$

Hence, choosing  $\eta = \varepsilon^2/8$ , we obtain that for all large enough integers  $n$ , with probability at least  $1 - 2\varepsilon$ ,  $\hat{G}_n(t) > \hat{G}_n(t_n^*)$  simultaneously for all  $t \in T_n$  with  $\|t - t_n^*\|_S = \varepsilon$ . Corollary 1 yields that for all large enough integers  $n$ , with probability at least  $1 - 2\varepsilon$ ,  $\|\hat{t}_n - t_n^*\|_S \leq \varepsilon$ . That is,  $\hat{t}_n - t_n^*$  converges in probability to 0.

*Conclusion.* We have proved that  $t_n^*$  converges in distribution to  $d^+ \pi_{\Theta - \theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); Z)$  for some Gaussian random variable  $Z$  and that  $\hat{t}_n - t_n^*$  converges to zero in probability, as  $n \rightarrow \infty$ . Hence, Slutsky's theorem implies the desired result.  $\square$

In the proof of Theorem 7, the convergence that we obtained in (10) actually holds in the  $L^2$  sense (see (9)). Therefore, Corollary 2 implies uniform convergence on all compact subsets in the  $L^2$  sense. Yet, it is not clear, from there, how to proceed and prove that  $\hat{t}_n - t_n^* \xrightarrow[n \rightarrow \infty]{} 0$  in  $L^2$ . Proving this convergence would yield an exact asymptotic quantification of the mean squared error of  $\hat{\theta}_n$ , since, it would yield that

$$n \mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2] \xrightarrow[n \rightarrow \infty]{} \mathbb{E}[\|d^+ \pi_{\Theta - \theta^*}^S(-S^{-1} \nabla \Phi(\theta^*); Z)\|^2]$$

where  $Z$  is a Gaussian vector as in the theorem. We leave the following question open:

OPEN QUESTION. *Is it true that under the assumptions of Theorem 7, for all large enough  $n$ ,  $\hat{\theta}_n$  has two moments, and that*

$$n\mathbb{E}[\|\hat{\theta}_n - \theta^*\|^2] \xrightarrow{n \rightarrow \infty} \mathbb{E}[\|d^+ \pi_{\Theta-\theta^*}^S(-S^{-1}\nabla\Phi(\theta^*); Z)\|^2]?$$

## 5. EXTENSION: CONVEX $U$ -ESTIMATION

The previous theory can be easily extended to more general convex empirical risks, e.g., when  $\Phi_n(\theta)$  is a  $U$ -statistic. With the same notation as in the previous sections, fix some positive integer  $k$  and let  $\phi : E^k \times \Theta_0 \rightarrow \mathbb{R}$  be symmetric and measurable in its first  $k$  arguments and convex in its last. Also assume that for all  $\theta \in \Theta_0$ ,  $\phi(\cdot, \theta) \in L^1(P^{\otimes k})$ , that is,  $\phi(X_1, \dots, X_k, \theta)$  is integrable. Set  $\Phi(\theta) = \mathbb{E}[\phi(X_1, \dots, X_k, \theta)]$  and, for all  $n \geq k$ ,

$$\Phi_n(\theta) = \frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < \dots < i_k \leq n} \phi(X_{i_1}, \dots, X_{i_k}, \theta).$$

Estimators obtained by minimizing such empirical risks are called  $U$ -estimators. Some relevant examples include:

1. Location estimators through depth functions: Let  $E = \Theta_0 = \Theta = \mathbb{R}^d$ ,  $k = d$  and  $\phi(x_1, \dots, x_d, \theta)$  be the volume of the  $d$ -dimensional simplex spanned by  $x_1, \dots, x_d, \theta$ , for all  $x_1, \dots, x_d, \theta \in \mathbb{R}^d$ . The minimizers of  $\Phi$  are then called Oja's population medians [44]. Note that  $\phi(x_1, \dots, x_d, \theta)$  is the absolute value of an affine function of  $\theta$ , hence, it is convex in  $\theta$ . We recover consistency and asymptotic normality of Oja's empirical medians (see [45]) as particular cases of our asymptotic theorems (see below for  $U$ -estimators). More generally, we refer to [58] for other definitions of medians that are  $U$ -estimators associated with depth functions.
2. Let  $E = \mathbb{R}$  and  $\Theta \subseteq \Theta_0 = \mathbb{R}$  and  $k \geq 1$ . [37] proposes a version of the median of mean estimator defined as a  $U$ -estimator obtained by computing an empirical median of all empirical averages of the form  $\frac{1}{k} \sum_{i \in I} X_i$ , for  $I \subseteq \{1, \dots, n\}$  of size  $k$ . That is,  $\phi(x_1, \dots, x_k, \theta) = \left| \frac{x_1 + \dots + x_k}{k} - \theta \right|$ , for all  $x_1, \dots, x_k, \theta \in \mathbb{R}$ . The difference with standard median of mean estimators [32, 33, 39] is that in [37], all possible subsamples of size  $k$ , with overlaps, are considered. Other frameworks, such as geometric medians of means in multivariate settings [36] can be considered as well. Note that in [37], the order  $k$  of the  $U$ -process is allowed to grow with the sample size  $n$  - we do not consider this setup here and leave it for future work.
3. More generally, aggregation of estimators that are based on overlapping subsamples, e.g., random forests [9] or bagging [8], which have attracted lots of interest in modern machine learning.
4. Scatter estimation and robustness: Let  $E = \mathbb{R}$ ,  $\Theta_0 = \mathbb{R}$ ,  $k = 2$  and  $\phi(x_1, x_2, \theta) = \ell(|x_1 - x_2|^p - \theta)$  where  $p \geq 1$  and  $\ell : \mathbb{R} \rightarrow \mathbb{R}$  is a convex function. When  $p = 2$  and  $\ell(u) = u^2$ ,  $u \in \mathbb{R}$ ,  $\hat{\theta}_n$  is simply twice the empirical variance of  $X_1, \dots, X_n$  and if  $\ell = h_c$  for some  $c > 0$  (recall the definition of  $h_c$  from Section 1.1), we obtain a robust version of the empirical variance. If now  $p = 1$  and  $\ell(u) = u^2$ ,  $u \in \mathbb{R}$ , we obtain Gini's mean absolute difference, while if  $\ell = |\cdot|$ , we obtain a proxy to a median absolute deviation (and intermediate robust versions if  $\ell = h_c$  for some  $c > 0$ ). In higher dimensions, one recovers the empirical covariance matrix of  $X_1, \dots, X_n$  by setting  $\phi(x_1, x_2, \theta) = \text{tr}(((x_1 - x_2)(x_1 - x_2)^\top - \theta)^2)$ , for all  $\theta \in \mathbb{R}^{d \times d} \approx \mathbb{R}^{d^2}$  and  $x_1, x_2 \in \mathbb{R}^d$ . Robust versions can be defined by taking the square root of the above, or applying Huber's loss  $h_c$  for some  $c > 0$ .



5. Empirical risk minimization where the choice of loss function itself depends on the data (e.g., for data driven procedures), see, e.g., [53].

Note that  $U$ -statistics depending on a parameter (here,  $\Phi_n(\theta), \theta \in \Theta_0$ ) have been studied as  $U$ -processes, see, e.g., [4, 41, 42]. Here, we first recall the classical law of large numbers and central limit theorem for  $U$ -statistics.

**THEOREM 8.** *Law of large numbers for  $U$ -statistics [20, Theorem 8.6] Let  $h : E^k \rightarrow \mathbb{R}^d$  be a symmetric, measurable map satisfying  $h \in L^1(P^{\otimes k})$ . Then,*

$$\frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < \dots < i_k \leq n} h(X_{i_1}, \dots, X_{i_k}) \xrightarrow{n \rightarrow \infty} \mathbb{E}[h(X_1, \dots, X_k)]$$

*almost surely.*

**THEOREM 9.** *Central limit theorem for multivariate  $U$ -statistics [22, Theorem 7.1], [20, Theorem 8.9] Let  $h : E^k \rightarrow \mathbb{R}^d$  be a symmetric, measurable map satisfying  $h \in L^2(P^{\otimes k})$ . Let  $\Sigma$  be the covariance matrix of  $\mathbb{E}[h(X_1, \dots, X_k) | X_1]$ <sup>6</sup>. For all  $n \geq k$ , let  $U_n = \frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < \dots < i_k \leq n} h(X_{i_1}, \dots, X_{i_k})$ . Then,*

$$\sqrt{n}(U_n - \mathbb{E}[h(X_1, \dots, X_k)]) \xrightarrow{n \rightarrow \infty} \mathcal{N}_d(0, k^2 \Sigma)$$

*in distribution.*

Theorem 4 obviously remains true in the context of  $U$ -estimation with convex loss. Proposition 1, Theorems 5 and 6 require more care but also remain true in this context. Proofs are deferred to Section D. Below, we rewrite Theorem 7 for  $U$ -estimators, where an extra multiplicative factor  $k$  appears in the limit, accounting for the dependence of the terms in the new definition of  $\Phi_n$ .

**THEOREM 10.** *Asymptotic distribution for  $U$ -estimators Let  $g : E^k \times \Theta_0 \rightarrow \mathbb{R}^d$  be a measurable selection of subgradients of  $\phi$ . Assume the following:*

- (i)  $\Phi$  has a unique minimizer  $\theta^*$  in  $\Theta$ , it is twice differentiable at  $\theta^*$  and  $S := \nabla^2 \Phi(\theta^*)$  is positive definite;
- (ii)  $g(\cdot, \theta^*) \in L^2(P^{\otimes k})$ ;
- (iii)  $\pi_{\Theta - \theta^*}^S$  has directional derivatives at  $-S^{-1} \nabla \Phi(\theta^*)$ .

*Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{n \rightarrow \infty} k d^+ \pi_{\Theta - \theta^*}^S(S^{-1} \nabla \Phi(\theta^*); Z)$$

*in distribution, where  $Z \sim \mathcal{N}_d(0, S^{-1} B S^{-1})$  and  $B = \text{var}(\mathbb{E}[g(X_1, \dots, X_k, \theta^*) | X_1])$ .*

Note the extra  $k$  factor in the limit in distribution.

---

<sup>6</sup> $\Sigma$  can also be written as  $\mathbb{E}[h(X_1, X_2, \dots, X_k) h(X_1, X'_2, \dots, X'_k)^\top] - \mathbb{E}[h(X_1, \dots, X_k)] \mathbb{E}[h(X_1, \dots, X_k)]^\top$ , that is, the covariance of the random vectors  $h(X_1, X_2, \dots, X_k)$  and  $h(X_1, X'_2, \dots, X'_k)$ , where  $X'_2, \dots, X'_k$  are such that  $X_1, X_2, \dots, X_k, X'_2, \dots, X'_k$  are iid.

## 6. CONCLUSION AND FUTURE DIRECTIONS

We have established the asymptotic properties of constrained  $M$ -estimators with a convex loss and a convex set of constraints, under minimal assumptions. In this work, asymptotics are only relative to the sample size  $n$ , while the dimension  $d$  is kept fixed.

In large dimensional problems, asymptotic theory can be approached from different angles. First, one may look at asymptotic distributions of low-dimensional projections of the  $M$ -estimator. For instance, in the context of linear regression, [6] proves the asymptotic normality of single coordinates of penalized  $M$ -estimators when the ratio  $d/n$  goes to some fixed, positive constant. A second angle consists of looking at the full, joint distribution of (a rescaled version of) the  $M$ -estimator  $\hat{\theta}_n$ , and prove that, for some distribution  $Q_d$  in  $\mathbb{R}^d$ , some specified distance (e.g., an integral probability metric) between the distribution of  $\hat{\theta}_n$  and  $Q_d$  goes to 0 as  $n, d \rightarrow \infty$  in a certain manner. When  $\hat{\theta}_n$  is simply the sample mean of  $X_1, \dots, X_n$ , such an approach has been studied and called *high dimensional central limit theorems* [12, 15]. However, to the best of our knowledge, such results do not exist for other  $M$ -estimators, even with convex loss.

In the context of  $U$ -estimators, we have also let the order  $k$  of the  $U$ -process be fixed. However, it may be relevant to also let  $k$  grow with the sample size (e.g., for median-of-means procedures). While the asymptotics of  $U$ -statistics with increasing order have been studied only recently [14], we leave this direction for future work on  $U$ -estimation.

## REFERENCES

- [1] Aleksandr D. Aleksandrov. Almost everywhere existence of the second differential of a convex function and some properties of convex functions. *Leningrad Univ. Ann.*, 37:3–35, 1939.
- [2] Jason Altschuler, Sinho Chewi, Patrik R. Gerber, and Austin Stromme. Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems*, 34:22132–22145, 2021.
- [3] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.
- [4] Miguel A. Arcones and Evarist Giné. Limit theorems for  $u$ -processes. *The Annals of Probability*, pages 1494–1542, 1993.
- [5] Pierre C Bellec and Takuya Koriyama. Asymptotics of resampling without replacement in robust and logistic regression. *arXiv preprint arXiv:2404.02070*, 2024.
- [6] Pierre C. Bellec, Yiwei Shen, and Cun-Hui Zhang. Asymptotic normality of robust  $M$ -estimators with convex penalty. *Electronic Journal of Statistics*, 16(2):5591–5622, 2022.
- [7] Rajendra Bhatia and John Holbrook. Riemannian geometry and matrix geometric means. *Linear algebra and its applications*, 413(2-3):594–618, 2006.
- [8] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- [9] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [10] Efim M. Bronshteyn and Leonid D. Ivanov. The approximation of convex sets by polyhedra. *Siberian Mathematical Journal*, 16(5):852–853, 1975.
- [11] Sourav Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, pages 2340–2381, 2014.
- [12] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.
- [13] Yuan Shih Chow and Henry Teicher. *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media, 2003.

- [14] Cyrus DiCiccio and Joseph Romano. CLT for  $U$ -statistics with growing dimension. *Statistica Sinica*, 32(1):323–344, 2022.
- [15] Xiao Fang and Yuta Koike. High-dimensional central limit theorems by Stein’s method. *The Annals of Applied Probability*, 31(4):1660–1686, 2021.
- [16] Herbert Federer. Geometric measure theory. *Springer*, 1969.
- [17] Luisa Turrin Fernholz. Von Mises calculus for statistical functionals. *Lecture Notes in Statistics*, 1983.
- [18] Charles J. Geyer. On the asymptotics of constrained  $M$ -estimation. *The Annals of statistics*, pages 1993–2010, 1994.
- [19] Shelby J. Haberman. Concavity and estimation. *The Annals of Statistics*, pages 1631–1661, 1989.
- [20] Norbert Henze. *Asymptotic Stochastics*, volume 10. Springer, 2024.
- [21] Charles J. Himmelberg, Thiruvengkatachari Parthasarathy, and F.S. Van Vleck. On measurable relations. *Fundamenta mathematicae*, 61:161–167, 1982.
- [22] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pages 308–334. Springer, 1992.
- [23] Richard B. Holmes. Smoothness of certain metric projections on Hilbert space. *Transactions of the American Mathematical Society*, 184:87–100, 1973.
- [24] Jean Honorio and Tommi Jaakkola. A unified framework for consistency of regularized loss minimizers. In *International Conference on Machine Learning*, pages 136–144. PMLR, 2014.
- [25] Olav Kallenberg. *Foundations of modern probability*. Springer, 1997.
- [26] Keith Knight. Asymptotic theory for  $M$ -estimators of boundaries. In *The Art of Semiparametrics*, pages 1–21. Springer, 2006.
- [27] Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.
- [28] V Koltchinskii. Bahadur-Kiefer approximation for spatial quantiles. In *Probability in Banach Spaces, 9*, pages 401–415. Springer, 1994.
- [29] Takuya Koriyama, Pratik Patil, Jin-Hong Du, Kai Tan, and Pierre C. Bellec. Precise asymptotics of bagging regularized  $M$ -estimators. *arXiv preprint arXiv:2409.15252*, 2024.
- [30] Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. Statistical inference for Bures-Wasserstein barycenters. *The Annals of Applied Probability*, 31(3):1264–1298, 2021.
- [31] Lucien Le Cam. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41(3):802–828, 1970.
- [32] Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2):906–931, 2020.
- [33] Matthieu Lerasle and Roberto I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.
- [34] Jessie Li. Inference for constrained extremum estimators. *Unpublished manuscript*, 2024.
- [35] Yen-Huan Li, Ya-Ping Hsieh, Nissim Zerbib, and Volkan Cevher. A geometric view on constrained  $M$ -estimators. *arXiv preprint arXiv:1506.08163*, 2015.
- [36] Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308, 2015.
- [37] Stanislav Minsker.  $U$ -statistics of growing order and sub-Gaussian mean estimators with sharp constants. *Mathematical statistics and learning*, 7(1):1–39, 2023.
- [38] Ilya Molchanov. *Theory of random sets*. Springer, 2005.
- [39] Arkadij S. Nemirovskij and David B. Yudin. Problem complexity and method efficiency in optimization. 1983.
- [40] Wojciech Niemiro. Asymptotics for  $M$ -estimators defined by convex minimization. *The Annals*

- of Statistics*, pages 1514–1533, 1992.
- [41] Deborah Nolan and David Pollard. U-processes: rates of convergence. *The Annals of Statistics*, pages 780–799, 1987.
  - [42] Deborah Nolan and David Pollard. Functional limit theorems for  $U$ -processes. *The Annals of Probability*, 16(3):1291–1298, 1988.
  - [43] Dominikus Noll. Directional differentiability of the metric projection in Hilbert space. *Pacific Journal of Mathematics*, 170(2):567–592, 1995.
  - [44] Hannu Oja. Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1(6):327–332, 1983.
  - [45] Hannu Oja and Ahti Niinimaa. Asymptotic properties of the generalized median in the case of multivariate normality. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 372–377, 1985.
  - [46] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.
  - [47] David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.
  - [48] R. Tyrrell Rockafellar. *Convex analysis*, volume 18. Princeton university press, 1970.
  - [49] Rolf Schneider. *Convex bodies: the Brunn–Minkowski theory*. Number 151. Cambridge university press, 2014.
  - [50] Alexander Shapiro. On differentiability of metric projections in  $\mathbb{R}^n$ . I. Boundary case. *Proceedings of the American Mathematical Society*, 99(1):123–128, 1987.
  - [51] Alexander Shapiro. Directionally nondifferentiable metric projection. *Journal of optimization theory and applications*, 81(1):203–204, 1994.
  - [52] Alexander Shapiro. Differentiability properties of metric projections onto convex sets. *Journal of Optimization Theory and Applications*, 169(3):953–964, 2016.
  - [53] Robert P. Sherman. U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric theory*, 10(2):372–395, 1994.
  - [54] Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011.
  - [55] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
  - [56] Aad W Van Der Vaart and Jon A Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996.
  - [57] Eduardo H. Zarantonello. Projections on convex sets in Hilbert space and spectral theory: Part I. Projections on convex sets: Part II. Spectral theory. In *Contributions to Nonlinear Functional Analysis*, pages 237–424. Academic Press, 1971.
  - [58] Yijun Zuo and Robert Serfling. General notions of statistical depth function. *Annals of statistics*, pages 461–482, 2000.

## APPENDIX A: ON CONVEX FUNCTIONS AND THEIR SUBGRADIENTS

In this section, we gather useful properties on subgradients of convex functions. Most of these properties are classical and we include their proofs for completeness.

LEMMA 4. *Let  $G_0 \subseteq \mathbb{R}^d$  be a convex set with non-empty interior and  $f : G_0 \rightarrow \mathbb{R}$  be a convex function. Let  $t_0 \in \text{int}(G_0)$  and  $\varepsilon > 0$  be such that  $B(t_0, \varepsilon) \subseteq \text{int}(G_0)$ . Then, for all vectors  $u \in \mathbb{R}^d$ ,  $u \in \partial f(t_0)$  if and only if  $f(t) \geq f(t_0) + u^\top(t - t_0)$ , for all  $t \in B(t_0, \varepsilon)$ .*

PROOF. The left-right implication trivially follows the definition of subgradients. Assume now that a vector  $u \in \mathbb{R}^d$  satisfies the right property. Fix  $t \in G_0$  be arbitrary and let us show that

$f(t) \geq f(t_0) + u^\top(t - t_0)$ . This is clear by assumption if  $t \in B(t_0, \varepsilon)$ , so let us assume that  $t \notin B(t_0, \varepsilon)$ . Let  $\lambda = \varepsilon / \|t - t_0\| \in (0, 1)$  and  $t_\lambda := t_0 + \lambda(t - t_0) \in B(t_0, \varepsilon)$ . Then, by assumption,  $f(t_\lambda) \geq f(t_0) + u^\top(t_\lambda - t_0) = f(t_0) + \lambda u^\top(t - t_0)$ . Moreover, by convexity of  $h$ ,  $f(t_\lambda) \leq (1 - \lambda)f(t_0) + \lambda f(t)$ . Rearranging yields the desired inequality.  $\square$

LEMMA 5. *Let  $G_0 \subseteq \mathbb{R}^d$  be a convex set and  $f : G_0 \rightarrow \mathbb{R}$  be any convex function. Then, for all  $t_0 \in G_0$ ,  $\partial f(t_0)$  is closed. Moreover, if  $t_0 \in \text{int}(G_0)$ , then,  $\partial f(t_0)$  is non-empty and compact.*

PROOF.

$\partial f(t_0)$  is closed. Fix  $t_0 \in G_0$  and let  $(u_n)_{n \geq 1}$  be a sequence of subgradients of  $h$  at  $t_0$ , assumed to converge to some  $u \in \mathbb{R}^d$ . For all  $t \in G_0$  and for all  $n \geq 1$ ,  $f(t) \geq f(t_0) + u_n^\top(t - t_0)$ . Taking the limit as  $n \rightarrow \infty$  shows that  $u \in \partial f(t_0)$ . Hence,  $\partial f(t_0)$  is closed.

$\partial f(t_0)$  is nonempty if  $t_0 \in \text{int}(G_0)$ . Let  $F = \{(t, y) \in G_0 \times \mathbb{R} : f(t) \leq y\}$  be the epigraph of  $h$ , which is a convex subset of  $\mathbb{R}^{d+1}$ . Let  $t_0 \in \text{int}(G_0)$ . The point  $(t_0, f(t_0))$  is a boundary point of  $F$  since  $(t_0, f(t_0) - \varepsilon) \notin F$  for all  $\varepsilon > 0$ . Let  $H \subseteq \mathbb{R}^{d+1}$  be a supporting hyperplane of  $F$  at this point and let  $v = (v_1, v_2)$  be a (non-zero) outward pointing normal vector to  $H$ , where  $v_1 \in \mathbb{R}^d$  and  $v_2 \in \mathbb{R}$ . This simply means that for all  $(t, y) \in F$ , the scalar product between  $v$  and  $(t, y) - (t_0, f(t_0))$  is non-positive. That is,

$$(12) \quad v_1^\top(t - t_0) + v_2(y - f(t_0)) \leq 0$$

for all  $t \in G_0$  and  $y \geq f(t)$ .

Let us show that necessarily,  $v_2 < 0$ . First, assume, for the sake of contradiction, that  $v_2 = 0$ . Then,  $v_1 \neq 0$ , because we have assumed that  $v = (v_1, v_2)$  is non-zero. Since  $t_0 \in \text{int}(G_0)$ , there exists  $t \in G_0$  such that  $v_1^\top(t - t_0) > 0$  (take  $t = t_0 + \varepsilon v_1$  for any small enough  $\varepsilon > 0$ ), which contradicts (12). Hence,  $v_2 \neq 0$ . Now, fixing any  $t \in G_0$  and taking  $y \rightarrow \infty$  in (12) shows that  $v_2$  must be negative. Now, taking  $y = f(t)$  in (12) yields, for all  $t \in G_0$ ,

$$f(t) \geq f(t_0) - v_2^{-1} v_1^\top(t - t_0).$$

That is,  $-v_2^{-1} v_1$  is a subgradient of  $h$  at  $t_0$ , so  $\partial f(t_0) \neq \emptyset$ .

$\partial f(t_0)$  is compact. It is now enough to check that for  $t_0 \in \text{int}(G_0)$ ,  $\partial f(t_0)$  is bounded. Fix  $\varepsilon > 0$  such that  $B(t_0, \varepsilon) \subseteq \text{int}(G_0)$ . Since  $h$  is continuous on  $\text{int}(G_0)$ , it is bounded on the compact set  $B(t_0, \varepsilon)$ . Let  $M := \max_{t \in B(t_0, \varepsilon)} f(t)$ . Let  $u \in \partial f(t_0)$  and assume that  $u \neq 0$ . Then, letting  $t = t_0 + \varepsilon u / \|u\| \in B(t_0, \varepsilon)$ , the definition of subgradients yields that  $M - f(t_0) \geq f(t) - f(t_0) \geq u^\top(t - t_0) = \varepsilon \|u\|$ . Hence,  $\partial f(t_0) \subseteq B(0, (M - f(t_0))/\varepsilon)$ .  $\square$

If  $t$  is a boundary point of  $G_0$ , then  $\partial f(t)$  might be empty. This is the case, for instance, for  $G_0 = \mathbb{R}^+$  and  $f : t \in \mathbb{R}^+ \mapsto -\sqrt{t}$ , which does not have any subgradient at 0.

LEMMA 6. *Let  $G_0 \subseteq \mathbb{R}^d$  be a convex set and  $f : G_0 \rightarrow \mathbb{R}$  be a convex function. Let  $t_0 \in \text{int}(G_0)$  and assume that  $h$  is differentiable at  $t_0$ . Then,  $\partial f(t_0) = \{\nabla f(t_0)\}$ .*

That is, if  $f$  is differentiable at some interior point of its domain, then its gradient is the only subgradient at that point. This property does not hold if  $t_0$  is a boundary point. For instance, let  $f : t \in \mathbb{R}^+ \mapsto 0$ , which is convex. Then, while it is differentiable at 0,  $\partial f(0) = \mathbb{R}^+$ .

PROOF. Let  $u \in \partial f(t_0)$ , where  $t_0 \in \text{int}(G_0)$ . Then, for all  $v \in \mathbb{R}^d$  and all small enough  $\varepsilon > 0$ ,

$$f(t_0 + \varepsilon v) - f(t_0) \geq \varepsilon u^\top v.$$

Dividing by  $\varepsilon$  and taking the limit as  $\varepsilon \rightarrow 0$  yields

$$\nabla f(t_0)^\top v \geq u^\top v.$$

Since this must hold for any  $v \in \mathbb{R}^d$ , one readily obtains that  $u = \nabla f(t_0)$ .  $\square$

LEMMA 7. *Let  $G_0 \subseteq \mathbb{R}^d$  be a convex set and  $f : G_0 \rightarrow \mathbb{R}$  be a convex function. Let  $t_0 \in G_0$  and assume that  $f$  is differentiable at  $t_0$ . Let  $(t_n)_{n \geq 1}$  be any sequence of points in  $G_0$  converging to  $t_0$ . For all  $n \geq 1$ , let  $u_n \in \partial f(t_n)$ . Then,  $u_n \xrightarrow{n \rightarrow \infty} \nabla f(t_0)$ .*

PROOF. Let  $\varepsilon > 0$  be such that  $B(t_0, \varepsilon) \subseteq \text{int}(G_0)$ . A similar argument as in the proof of compactness of  $\partial f(t_0)$  in Lemma 5 yields that  $\bigcup_{t \in B(t_0, \varepsilon)} \partial f(t)$  is bounded, so the sequence  $(u_n)_{n \geq 1}$  must be bounded. Therefore, it is sufficient to prove that any converging subsequence must converge to  $\nabla f(t_0)$ . Since we could simply relabel the indices of the sequence, let us simply assume that  $u_n \xrightarrow{n \rightarrow \infty} u$  for some  $u \in \mathbb{R}^d$ . For all  $t \in G_0$  and all  $n \geq 1$ ,

$$f(t) \geq f(t_n) + u_n^\top (t - t_n).$$

Recall that  $f$  is continuous on  $\text{int}(G_0)$ , so taking the limit as  $n \rightarrow \infty$  in the previous display gives

$$f(t) \geq f(t_0) + u^\top (t - t_0),$$

so  $u \in \partial f(t_0)$ . Lemma 6 implies that  $u = \nabla f(t_0)$ .  $\square$

The following lemma is more general and will allow to connect subgradients and directional derivatives.

LEMMA 8. *Let  $G_0 \subseteq \mathbb{R}^d$  be a convex set and  $f : G_0 \rightarrow \mathbb{R}$  be a convex function. Let  $x \in \text{int}(G_0)$  and let  $(x_n)_{n \geq 1}$  be a sequence of points in  $\text{int}(G_0)$  converging to  $x$ . For each  $n \geq 1$ , let  $u_n \in \partial f(x_n)$ . Then, the sequence  $(u_n)_{n \geq 1}$  is bounded and any of its converging subsequences converges to some element of  $\partial f(x)$ .*

PROOF. Let  $\varepsilon > 0$  satisfying  $B(x, \varepsilon) \subseteq \text{int}(G_0)$ . Without loss of generality, let us assume that  $x_n \in B(x, \varepsilon)$  for all  $n \geq 1$ . Convexity of  $f$  yields that it is locally Lipschitz [49, Theorem 1.5.3], and hence, there is some  $L > 0$  such that  $\|u\| \leq L$  for all  $u \in \partial f(y), y \in B(x, \varepsilon)$ . Hence,  $(u_n)_{n \geq 1}$  is bounded.

Now, consider a converging subsequence of  $(u_n)_{n \geq 1}$  which, up to renumbering, we still denote by  $(u_n)_{n \geq 1}$ . Let  $u$  be its limit. Then, for all  $y \in G_0$  and  $n \geq 1$ ,

$$f(y) \geq f(x_n) + u_n^\top (y - x_n).$$

Since  $f$  is continuous at  $x$ , all terms have a limit as  $n \rightarrow \infty$  and we obtain, for all  $y \in G_0$ ,

$$f(y) \geq f(x) + u^\top (y - x).$$

That is,  $u \in \partial f(x)$ .  $\square$



LEMMA 9. *Let  $G_0 \subseteq \mathbb{R}^d$  be a convex set and  $f : G_0 \rightarrow \mathbb{R}$  be a convex function. Let  $x \in \text{int}(G_0)$ . Then, for all  $t \in \mathbb{R}^d$ ,*

$$d^+f(x; t) = h_{\partial f(x)}(t)$$

where we recall that  $h_{\partial f(x)}$  is the support function of  $\partial f(x)$ .

PROOF. Let us first check that for all  $u \in \partial f(x)$ ,  $u^\top t \leq d^+f(x; t)$ . To obtain this, note that by definition of  $u$ , we have that  $f(x + \varepsilon t) \geq f(x) + \varepsilon t^\top u$  for all  $\varepsilon > 0$  and, hence, by rearranging and taking the limit as  $\varepsilon \rightarrow 0$ ,  $d^+f(x; t) \geq u^\top t$ . Now, let us simply check the existence of  $u \in \partial f(x)$  satisfying  $u^\top t = d^+f(x; t)$ : This will end the proof.

For all large enough integers  $n$  (so  $x + t/n \in \text{int}(G)$ ), let  $u_n \in \partial f(x + t/n)$ . Then,  $f(x) \geq f(x + t/n) - u_n^\top t/n$  which, after rearranging, gives:

$$n(f(x + t/n) - f(x)) \leq u_n^\top t.$$

The left hand side goes to  $d^+f(x; t)$  and, by Lemma 8, the right hand side has a subsequence that goes to  $u^\top t$  for some  $u \in \partial f(x)$ . We thus obtain that  $d^+f(x; t) \leq u^\top t$ , which is what we aimed for.  $\square$

LEMMA 10 (First order condition). *Let  $G_0$  be an open convex set and  $G \subseteq G_0$  be closed and convex. Let  $f : G_0 \rightarrow \mathbb{R}$  be a convex function and  $x_* \in G$ . Let  $C$  be the support cone to  $G$  at  $x_*$ . Then,*

$$f(x) \geq f(x_*), \forall x \in G \iff h_{\partial f(x_*)}(t) \geq 0, \forall t \in C.$$

In particular, we recover the standard first order condition if  $f$  is differentiable at  $x^*$ , that is,  $x^*$  is a minimizer of  $f$  on  $G$  if and only if  $\nabla f(x^*)^\top t \geq 0$  for all  $t \in C$ .

PROOF. Let  $T$  be the tangent cone to  $G$  at  $x^*$ , so  $C$  is the closure of  $T$ . It is clear that  $x^*$  is a minimizer of  $f$  on  $G$  if and only if  $d^+f(x^*; t) \geq 0$ , that is,  $h_{\partial f(x^*)}(t) \geq 0$  by Lemma 9. The result follows from the continuity of  $h_{\partial f(x^*)}$ .  $\square$

LEMMA 11 (Monotonicity of subgradients). *Let  $G_0 \subseteq \mathbb{R}^d$  be a convex set and  $f : G_0 \rightarrow \mathbb{R}$  be a convex function. Then, for all  $t_1, t_2 \in G_0$  and  $u_1 \in \partial f(t_1), u_2 \in \partial f(t_2)$ , we have  $(t_1 - t_2)^\top (u_1 - u_2) \geq 0$ .*

For differentiable, convex functions on  $\mathbb{R}$ , this lemma simply says that the derivative is non-decreasing.

PROOF. By definition of subgradients,

$$f(t_1) \geq f(t_0) + u_0^\top (t_1 - t_0)$$

and

$$f(t_0) \geq f(t_1) + u_1^\top (t_0 - t_1).$$

Adding these two inequalities yields the result.  $\square$

LEMMA 12. *Let  $f$  be a random convex function defined on a convex set  $G_0 \subseteq \mathbb{R}^d$  and let  $x_0 \in \text{int}(G_0)$ . Assume that for all  $x \in G_0$ ,  $f(x)$  is integrable and let  $F(x) = \mathbb{E}[f(x)]$ . Then, for all  $t \in \mathbb{R}^d$ ,*

$$\mathbb{E}[d^+f(x_0; t)] = d^+F(x_0; t).$$

In particular, if  $F$  is differentiable at  $x_0$ , so is  $f$  almost surely.

PROOF. Fix  $t \in \mathbb{R}^d$ . For simplicity (and without loss of generality), assume that  $x_0 - t, x_0 + t \in G_0$ . First, we have that  $d^+ f(x_0; t) = \lim_{n \rightarrow \infty} n(f(x_0 + t/n) - f(x_0))$  almost surely. Moreover, convexity of  $f$  yields that:

- $n(f(x_0 + t/n) - f(x_0))$  is non-increasing with  $n$ ;
- $f(x_0) - f(x_0 - t) \leq n(f(x_0 + t/n) - f(x_0)) \leq f(x_0 + t) - f(x_0)$

where both bounds in the last display are integrable. Therefore, monotone convergence implies that

$$\mathbb{E}[d^+ f(x_0; t)] = \lim_{n \rightarrow \infty} \mathbb{E}[n(f(x_0 + t/n) - f(x_0))] = \lim_{n \rightarrow \infty} n(F(x_0 + t/n) - F(x_0)) = d^+ F(x_0; t).$$

Now, let us assume that  $F$  is differentiable at  $x_0$ . Since  $f$  is convex, in order to show that it is almost surely differentiable at  $x_0$ , it is enough to show that it has partial derivatives at  $x_0$  along all canonical basis directions with probability 1 (see [49, Theorem 1.5.8]). That is, we need to show that with probability 1, for all canonical basis vectors  $e$ , it holds that  $d^+ f(x_0; e) = d^+ f(x_0, -e)$ . Convexity of  $f$  yields that the right hand side is larger or equal to the left hand side, and the first part of this lemma implies that the expected difference is zero. Hence, both sides are equal with probability 1, which concludes the proof.  $\square$

REMARK 7. *In the previous lemma, convexity of the random function  $f$  is key. For instance, let  $f_0(\theta) = |\theta|$  and  $f_1(\theta) = -|\theta|$ , for all  $\theta \in \mathbb{R}$  ( $d = 1$  here). Set  $f = f_I$ , where  $I$  is a Bernoulli random variable with  $P(I = 0) = P(I = 1) = 1/2$ . Then, with probability 1,  $f_I$  is not differentiable at 0. Yet,  $F$  is the constant function equal to 0, which is differentiable at 0.*

## APPENDIX B: ON METRIC PROJECTIONS

The following lemma is a very standard result on projections on closed, convex sets in Euclidean spaces. We choose to state it here with our notation for the ease of the reader.

LEMMA 13. *Let  $G \subseteq \mathbb{R}^d$  be a non-empty, closed, convex set and  $S \in \mathbb{R}^{d \times d}$  be symmetric, positive definite. Then, for all  $z \in \mathbb{R}^d$ ,  $\pi_G^S(z)$  is the unique  $z^* \in G$  satisfying*

$$\langle x - z^*, z - z^* \rangle_S \leq 0, \quad \forall x \in G.$$

*In particular,  $z - \pi_G^S(z)$  is in the normal cone to  $G$  at  $\pi_G^S(z)$  with respect to  $S$ . Moreover,  $\pi_G^S$  is non-expansive with respect to  $\|\cdot\|_S$ , that is, for all  $z, z' \in \mathbb{R}^d$ ,*

$$\|\pi_G^S(z) - \pi_G^S(z')\|_S \leq \|z - z'\|_S.$$

PROOF. Let  $z \in \mathbb{R}^d$ . Then, by definition of  $\pi_G^S(z)$ , we have, for all  $x \in G$ , that  $\|z - \pi_G^S(z)\|_S^2 \leq \|z - x\|_S^2$ . Expanding these Euclidean norms and rearranging yield that  $\pi_G^S(z)$  does satisfy the first inequality of the lemma. Now, assume that  $z^* \in G$  also satisfies this inequality. Reverse engineering simply implies that  $\|z - z^*\|_S^2 \leq \|z - x\|_S^2$  for all  $x \in G$ , and hence,  $z^* = \pi_G^S(z)$ .

Non-expansiveness of  $\pi_G^S$  is a direct consequence of the first inequality of the lemma. Indeed, it implies both that

$$\langle \pi_G^S(z') - \pi_G^S(z), z - \pi_G^S(z) \rangle_S \leq 0$$

and

$$\langle \pi_G^S(z) - \pi_G^S(z'), z' - \pi_G^S(z') \rangle_S \leq 0.$$

Summing these two inequalities yields that

$$(13) \quad \begin{aligned} \|\pi_G^S(z') - \pi_G^S(z)\|_S^2 &\leq \langle \pi_G^S(z') - \pi_G^S(z), z' - z \rangle_S \\ &\leq \|\pi_G^S(z') - \pi_G^S(z)\|_S \|z - z'\|_S \end{aligned}$$

by Cauchy-Schwarz inequality, which yields the result.  $\square$

### APPENDIX C: ON THE DIRECTIONAL DIFFERENTIABILITY OF METRIC PROJECTIONS

Here, we gather several facts on the existence of directional derivatives of metric projections and their formulas. For simplicity, we choose to state and prove all the results of this section for the standard, canonical Euclidean structure of  $\mathbb{R}^d$ , that is, for  $S = I_d$ . All the results and formulas extend in a straightforward manner to general symmetric, positive definite  $S$ .

Let  $G \subseteq \mathbb{R}^d$  be a non-empty, closed, convex set. Although the map  $\pi_G$  is non-expansive, it does not necessarily have directional derivatives at any point. A counterexample (among others) is given in [51] and can be described easily for  $d = 2$ : Let  $G$  be the convex hull of all points of the form  $(\cos(\pm\pi/2^k), \sin(\pm\pi/2^k))$ , for  $k = 0, 1, 2, \dots$  and of  $(1, 0)$ . Then, letting  $x = (a, 0)$  for any  $a > 0$ ,  $\pi_G(x) = (1, 0)$  and  $\pi_G$  does not have a directional derivative at  $(0, \pm 1)$ . Roughly speaking, this stems from the fact that the boundary of  $G$  is not twice directionally differentiable at  $(1, 0)$  in neither directions  $(0, 1)$  or  $(0, -1)$ .

First, it is obvious that if  $x \in \text{int}(G)$ , then  $\pi_G$  coincides with the identity map on a neighborhood of  $x$ , and hence, it is differentiable (and hence, it has directional derivatives) at  $x$ , and its Jacobian at  $x$  is the identity matrix. If  $x \in \partial G$  then  $\pi_G$  is not always differentiable at  $x$  but it has directional derivatives in all directions:

LEMMA 14. [50], [57, Lemma 4.6] *Let  $G \subseteq \mathbb{R}^d$  be non-empty, closed and convex and let  $x \in \partial G$ . Then,  $\pi_G$  has directional derivatives at  $x$ , given by*

$$d^+ \pi_G(x; \cdot) = \pi_C$$

where  $C$  is the support cone to  $G$  at  $x$ .

In particular,  $d^+ \pi_G$  is differentiable at  $x$  if and only if  $\pi_C$  is a linear map, that is,  $C$  is a linear subspace, if and only if  $G$  is included in a strict affine subspace of  $A \subseteq \mathbb{R}^d$  and  $x$  is in the relative interior of  $G$  (in that case,  $A = x + C$ ).

When  $x \in \mathbb{R}^d \setminus G$ , we have the following sufficient condition for differentiability of  $\pi_G$  at  $x$ .

LEMMA 15. [23, Lemma 1 and Theorem 2]

*Let  $G \subseteq \mathbb{R}^d$  be non-empty, closed and convex. Let  $x \in \mathbb{R}^d \setminus G$  and assume that the boundary of  $G$  is of class  $C_k$  in a neighborhood of  $\pi_G(x)$ , for some  $k \geq 2$ . Then,  $\pi_G$  is of class  $C^{k-1}$  in a neighborhood of  $x$ .*

*Further assume that  $\text{int}(G) \neq \emptyset$  and  $0 \in \text{int}(G)$  and let  $\rho_G$  be the gauge function of  $G$ , defined by  $\rho_G(y) = \inf\{\lambda > 0 : y \in \lambda G\}$ , for all  $y \in \mathbb{R}^d$ . Then,  $\rho_G$  is twice differentiable at  $y := \pi_G(x)$  with  $\nabla \rho_G(y) \neq 0$  and*

$$(14) \quad d\pi_G(x; \cdot) = \left( I_d + \frac{\|x - y\|}{\|\nabla \rho_G(y)\|} \pi_{(x-y)^\perp} \nabla^2 \rho_G(y) \right)^{-1} \pi_{(x-y)^\perp},$$

where we have identified linear maps with their matrices in the canonical basis.

Note that the assumption that  $0 \in \text{int}(G)$  is made with no loss of generality, since  $G$  could be replaced with  $G - y_0$  for some  $y_0 \in \text{int}(G)$  (and  $\rho_G$  would be replaced with  $\rho_{G-y_0}$  in (14)). Note also that under the assumptions of the lemma, for all  $z \in \mathbb{R}^d$ ,  $d\pi_G(x; z) \in (x - y)^\perp$ .

One can easily derive a simpler formula than (14) by identifying  $\mathbb{R}^d$  with  $\mathbb{R}^{d-1} \times \mathbb{R}$ ,  $y$  with  $(0, 0)$ ,  $x$  with  $(0, -t)$  for some  $t > 0$  and by locally identifying  $G$  with the epigraph of a twice differentiable convex map  $f : \mathbb{R}^{d-1} \rightarrow \mathbb{R}$  with  $f(0) = 0$  and  $\nabla f(0) = 0$ . Then, for all  $z = (z_1, z_2) \in \mathbb{R}^{d-1} \times \mathbb{R}$ ,  $d\pi_G(x; z) = ((I_{d-1} + t\nabla^2 f(0))^{-1}z_1, 0)$ .

A simple example to have in mind is that of  $G = B(0, R)$ . Then, for all  $x \in \mathbb{R}^d$  with  $\|x\| > R$ , we obtain

$$d\pi_G(x; z) = \frac{R}{\|x\|} \pi_{x^\perp}(z), \quad \forall z \in \mathbb{R}^d.$$

Therefore, in that case,  $d\pi_G(x; \cdot)$  is a rescaled version of the projection onto  $G$ , where the scaling factor depends on both the distance from  $x$  to  $G$  and the curvature of  $G$  at  $\pi_G(x)$ .

If  $G$  is defined by smooth, convex constraints, we have the following result which guarantees the existence of directional derivatives of  $\pi_G$ .

LEMMA 16. [52, Theorem 3.2] *Let  $g_1, \dots, g_p : \mathbb{R}^d \rightarrow \mathbb{R}$  be twice differentiable, convex functions and let  $G = \{x \in \mathbb{R}^d : g_j(x) \leq 0, j = 1, \dots, p\}$ . Assume Slater's qualification constraint: There exists  $y_0 \in \mathbb{R}^d$  with  $g_j(y_0) < 0$  for all  $j = 1, \dots, p$ . Then,  $\pi_G$  has directional derivatives everywhere.*

Let us now look at further cases where  $\partial G$  is not necessarily differentiable in a neighborhood of  $\pi_G(x)$ . First, let us explore the simple case where  $G$  is a closed, convex cone and  $\pi_G(x) = 0$ .

LEMMA 17. *Let  $C \subseteq \mathbb{R}^d$  be a non-empty, closed convex cone,  $S \in \mathbb{R}^{d \times d}$  be symmetric, positive definite and  $x \in \mathbb{R}^d$  satisfying  $x^\top y \leq 0$  for all  $y \in C$ . Then,  $\pi_C$  is directionally differentiable at  $x$  and for all  $z \in \mathbb{R}^d$ , the directional derivative of  $\pi_C$  at  $x$  in the direction  $z$  is given by*

$$d^+ \pi_C(x; z) = \lim_{\varepsilon \downarrow 0} \frac{\pi_C(x + \varepsilon z)}{\varepsilon} = \pi_{C_x}(z).$$

Recall the notation  $C_x = \{y \in C : x^\top y = 0\} = C \cap x^\perp$ . Note that, with the notation of the lemma, the assumption that  $x^\top y \leq 0$  for all  $y \in C$  (that is,  $x$  is in the polar of  $C$ ) implies that  $\pi_C(x) = 0$ .

PROOF. Let  $\varepsilon > 0$ . Since, for all  $y \in \mathbb{R}^d$ ,  $y \in C \iff \varepsilon y \in C$ , we have that

$$\begin{aligned} \pi_C(x + \varepsilon z) &= \operatorname{argmin}_{y \in C} \|x + \varepsilon z - y\|^2 \\ &= \varepsilon \operatorname{argmin}_{y \in C} \|x + \varepsilon z - \varepsilon y\|^2 \\ &= \varepsilon \operatorname{argmin}_{y \in C} \left( x^\top(z - y) + \frac{\varepsilon}{2} \|z - y\|^2 \right). \end{aligned}$$

If  $C \subseteq x^\perp$ , then  $x^\top y = 0$  for all  $y \in C$  so the previous display implies that  $\pi_C(x + \varepsilon z) = \varepsilon \pi_C(z)$ , yielding the desired result in that case, since  $C_x = C$ .

Let us now assume that  $C_x$  is a strict subset of  $C$ . That is, there are  $y \in C$  with  $x^\top y < 0$ . Our goal is still to show that  $y_\varepsilon := \operatorname{argmin}_{y \in C} \left( x^\top(z - y) + \frac{\varepsilon}{2} \|z - y\|^2 \right) \xrightarrow{t \rightarrow 0} \pi_{C_x}(z)$ .

First, note that for all  $\varepsilon > 0$ , this vector  $y_\varepsilon$  is well defined by strong convexity of the function that it minimizes and the fact that  $C$  is a closed convex set. Now, let  $t_\varepsilon = -x^\top y_\varepsilon$ . Then, by definition,  $y_\varepsilon \in C_{x, t_\varepsilon}$  for all  $\varepsilon > 0$ . Moreover, it is clear that

$$y_\varepsilon = \operatorname{argmin}_{y \in C_{x, t_\varepsilon}} \|z - y\|^2 = \pi_{C_{x, t_\varepsilon}}(z).$$

So, what we have to show is that  $\pi_{C_{x,t_\varepsilon}}(z) \xrightarrow{\varepsilon \rightarrow 0} \pi_{C_x}(z)$ .

First, let us check that  $t_\varepsilon \xrightarrow{\varepsilon \rightarrow 0} 0$ . The fact that  $t_\varepsilon \geq 0$  is clear from the fact that  $x^\top y \leq 0$  for all  $y \in C$  (and, in particular, for  $y = y_\varepsilon$ ). Moreover, for all  $\varepsilon > 0$ ,

$$x^\top z + t_\varepsilon = x^\top(z - y_\varepsilon) \leq x^\top(z - y_\varepsilon) + \frac{\varepsilon}{2} \|z - y_\varepsilon\|^2 \leq x^\top(z - y) + \frac{\varepsilon}{2} \|z - y\|^2$$

for all  $y \in C$ , by definition of  $y_\varepsilon$ . Choosing  $y = \pi_{C_x}(z)$  yields that  $t_\varepsilon \leq \frac{\varepsilon}{2} d(z, C_x)^2$ . Therefore,  $t_\varepsilon \xrightarrow{\varepsilon \rightarrow 0} 0$ .

Finally, in order to achieve our objective, it is sufficient to show that given any sequence  $(\varepsilon_n)_{n \geq 1}$  of positive number converging to 0,  $y_{\varepsilon_n} \xrightarrow{n \rightarrow \infty} \pi_{C_x}(z)$ . Consider such a sequence. For simplicity, let us denote by  $y_n := y_{\varepsilon_n}$ ,  $t_n := t_{\varepsilon_n}$  and  $C_n := C_{x,t_n}$ .

Let us start by showing that the sequence  $(y_n)_{n \geq 1}$  is bounded. As already mentioned above, since we have assumed that  $C_x$  is a strict subset of  $C$ , there must exist some  $y \in C$  with  $\alpha := -x^\top y > 0$ . For all  $n \geq 1$ , let  $\lambda_n = t_n/\alpha$ , so that  $\lambda_n y \in C_n$  for all  $n \geq 1$ . Therefore,  $\lambda_n y = \pi_{C_n}(\lambda_n y)$  and, since  $\pi_{C_n}$  is non-expansive (see Lemma 13), we have that

$$\begin{aligned} \|y_n\| &\leq \|y_n - \lambda_n y\| + \lambda_n \|y\| \\ &= \|\pi_{C_n}(z) - \pi_{C_n}(\lambda_n y)\| + \lambda_n \|y\| \\ &\leq \|z - \lambda_n y\| + \lambda_n \|y\| \\ &\leq \|z\| + 2\lambda_n \|y\| \\ &= \|z\| + 2\alpha^{-1} t_n \|y\| \end{aligned}$$

which is bounded since we have shown, earlier, that  $t_n \xrightarrow{n \rightarrow \infty} 0$ . The first and last inequalities above are simply the triangle inequality.

Now, since the sequence  $(y_n)_{n \geq 1}$  is bounded, in order to prove that it converges to  $\pi_{C_x}(z)$ , it is sufficient to check that any of its converging subsequences converges to that same point. Up to renumbering, let us simply assume that  $y_n \xrightarrow{n \rightarrow \infty} y^*$  for some  $y^* \in \mathbb{R}^d$ , and show that  $y^* = \pi_{C_x}(z)$ . Also, without loss of generality (since we could otherwise consider a further subsequence), let us assume that  $(t_n)_{n \geq 1}$  is decreasing. First, since  $y_n \in C$  for all  $n \geq 1$  and  $C$  is closed, it must hold that  $y^* \in C$ . Moreover, since  $-\langle x, y_n \rangle_S = t_n \xrightarrow{n \rightarrow \infty} 0$  it must hold that  $x^\top y^* = 0$ . Therefore,  $y^* \in C_x$ .

Hence, by Lemma 13, in order to check that  $y^* = \pi_{C_x}(z)$ , it is sufficient to show that for all  $y \in C_x$ ,  $(z - y^*)^\top (y - y^*) \leq 0$ . Let  $y \in C_x$  be arbitrary. Let  $(w_n)_{n \geq 1}$  be a sequence converging to  $y$  and such that  $w_n \in C_n$  for all  $n \geq 1$ . Such a sequence can be constructed, for instance, by taking  $w_n$  as the unique intersection of the affine hyperplane  $\{w \in \mathbb{R}^d = x^\top w = t_n\}$  with the segment connecting  $y_1$  and  $y$ . Then, since  $y_n = \pi_{C_n}(z)$ , Lemma 13 yields that  $(z - y_n)^\top (w_n - y_n) \leq 0$ , for all  $n \geq 1$ . Taking the limit as  $n \rightarrow \infty$  yields that  $(z - y^*)^\top (w - y^*) \leq 0$ . This concludes the proof.  $\square$

As a consequence of this lemma, we obtain the following result. A closed, convex set  $G$  is called locally conic at  $y \in G$  if and only if there exists  $r > 0$  such that  $G \cap B(y, r) = (y + C) \cap B(y, r)$  where  $C$  is the support cone to  $G$  at  $y$ .

**LEMMA 18.** *Let  $G \subseteq \mathbb{R}^d$  be a non-empty, closed, convex set and  $x \in \mathbb{R}^d$ . If  $G$  is locally conic at  $\pi_G(x)$ , then  $\pi_G$  has directional derivatives at  $x$  given by*

$$d^+ \pi_G(x; \cdot) = \pi_{C_u}$$

where  $C$  is the support cone to  $G$  at  $\pi_G(x)$  and  $u = x - \pi_G(x)$ .

PROOF. Set  $y = \pi_G(x)$ . Since we have, for all  $z \in \mathbb{R}^d$ ,  $\pi_G(z) = y + \pi_G(z - y)$  (this is easy to check using Lemma 13 for instance), one may simply assume that  $y = 0$ , without loss of generality.

Let  $r > 0$  be such that  $G \cap B(0, r) = C \cap B(0, r)$ . Let  $z \in \mathbb{R}^d$ . Since  $G$  is locally conic at  $y = 0$  and  $\pi_G$  is continuous (since it is non-expansive, see Lemma 13),  $\pi_G(x + \varepsilon z) \in G \cap B(0, r)$  for all small enough  $\varepsilon > 0$ . Hence,  $\pi_G(x + \varepsilon z) = \pi_{G \cap B(0, r)}(x + \varepsilon z) = \pi_{C \cap B(0, r)}(x + \varepsilon z)$ .

Now, again using Lemma 13, we have that  $\pi_C(x) = 0$ . Hence, again by continuity of  $\pi_C$ , it holds that for all small enough  $\varepsilon > 0$ ,  $\pi_C(x + \varepsilon z) \in B(0, r)$ , hence,  $\pi_C(x + \varepsilon z) = \pi_{C \cap B(0, r)}(x + \varepsilon z)$  for such small enough  $\varepsilon > 0$ .

Finally, we have obtained that for all small enough  $\varepsilon > 0$ ,

$$\pi_G(x + \varepsilon z) = \pi_C(x + \varepsilon z).$$

The conclusion follows using Lemma 17. □

An important case of locally conic convex sets is that of convex (possibly unbounded) polyhedra, that is, intersections of finitely many closed affine halfspaces. Indeed, we have the following lemma.

LEMMA 19. *All convex polyhedras are locally conic at any point.*

As a consequence of this lemma, for all closed, convex polyhedra  $G \subseteq \mathbb{R}^d$ ,  $\pi_G$  has directional derivatives everywhere and  $d^+ \pi_G(x; \cdot) = \pi_{C_{x-\pi_G(x)}}$  for all  $x \in \mathbb{R}^d$ , where  $C$  is the tangent cone (which coincides with the support cone) to  $G$  at  $x$ . Note that this is also a particular case of Lemma 16 above.

PROOF. Let  $G$  be a convex polyhedra and  $x \in \mathbb{R}^d$ . If  $x \notin G$ , the result is vacuous, since the tangent cone to  $G$  at  $x$  is empty, as well as  $G \cap B(0, r)$  for all small enough  $r > 0$ . If  $x \in \text{int}(G)$ , the result is also trivial, since in that case, the tangent cone to  $G$  at  $x$  is  $\mathbb{R}^d$ .

Now, let  $x \in \partial G$ . Write  $K = H_1 \cap \dots \cap H_p$  where  $H_1, \dots, H_p$  are closed affine halfspaces and  $p \geq 1$  is an integer. Without loss of generality (or else, simply reorder  $H_1, \dots, H_p$ ), assume that  $x \in \partial H_j$  for  $j = 1, \dots, r$  and  $x \notin \partial H_j$  for  $j = r + 1, \dots, p$ , for some  $r \in \{1, \dots, p\}$ . That is,  $H_1, \dots, H_r$  are exactly those halfspaces whose bounding hyperplane contains  $x$ . Let  $B = (H_1 - x) \cap \dots \cap (H_r - x)$ . This is a closed, convex cone, as the intersection of closed, convex cones. Our goal is to show that  $B$  coincides with  $C$ , the support cone to  $G$  at  $x$ . Indeed, then, it is easy to see that  $G \cap B(x, r) = (x + B) \cap B(0, r)$  for all small enough  $r > 0$ : It suffices to take any  $r \leq \min_{j \geq r+1} d(x, \partial H_j)$ .

For all  $y \in K$ ,  $y \in H_1 \cap \dots \cap H_p \subseteq H_1 \cap \dots \cap H_r$  so  $y - x \in (H_1 - x) \cap \dots \cap (H_r - x) = B$ . Hence,  $B$  contains  $G - x$  and, since  $B$  is a closed, convex cone, it also contains  $C$ . Conversely, let  $v \in B$  and let us show that  $x + \varepsilon v \in G$  for some small enough  $\varepsilon > 0$ . This will yield that  $v \in C$ . For all  $j = 1, \dots, r$ ,  $v \in H_j - x$  so  $x + v \in H_j$ . Now, by definition of  $r$ ,  $x \in \text{int}(H_j)$  for all  $j = r + 1, \dots, p$ , so there exists  $\varepsilon > 0$  such that  $x + \varepsilon v \in H_j$  for all  $j = r + 1, \dots, p$ . Therefore,  $x + \min(1, \varepsilon)v \in K$ , yielding that  $v \in C$  as desired. □

Even though, in general, when  $\pi_G$  has directional derivatives at some  $x \in \mathbb{R}^d$ , it does not necessarily hold that  $d^+ \pi_G(x; \cdot) = \pi_{C_{x-\pi_G(x)}}$ , where  $C$  is the support cone to  $G$  at  $\pi_G(x)$ , the following result holds true.

LEMMA 20. *Let  $G \subseteq \mathbb{R}^d$  be a non-empty, closed, convex set and  $x \in \mathbb{R}^d$ . Let  $C$  be the support cone to  $G$  at  $\pi_G(x)$ . Assume that  $\pi_G$  has directional derivatives at  $x$ . Then, for all  $z \in \mathbb{R}^d$ ,  $d^+ \pi_G(x; z) \in C_{x-\pi_G(x)}$ .*



PROOF. The fact that  $d^+\pi_G(x; z) \in C$  is clear from the facts that, for all  $\varepsilon > 0$ ,  $\varepsilon^{-1}(\pi_G(x + \varepsilon z) - \pi_G(x)) \in C$ , and  $C$  is closed. Hence, we only need to show that  $d^+\pi_G(x; z)$  is orthogonal to  $x - \pi_G(x)$ . For all  $\varepsilon > 0$ , Lemma 13 yields that

$$(x + \varepsilon z - \pi_G(x + \varepsilon z))^\top (\pi_G(x) - \pi_G(x + \varepsilon z)) \leq 0.$$

Using the fact that  $\pi_G(x + \varepsilon z) = \pi_G(x) + \varepsilon d^+\pi_G(x; z) + o(\varepsilon)$  as  $\varepsilon \rightarrow 0$ , we obtain

$$-\varepsilon(x - \pi_G(x))^\top d^+\pi_G(x; z) + o(\varepsilon) \leq 0,$$

hence, by dividing by  $\varepsilon$  and letting  $\varepsilon \rightarrow 0$ ,  $(x - \pi_G(x))^\top d^+\pi_G(x; z) \geq 0$ . Moreover, again by Lemma 13,  $(x - \pi_G(x))^\top d^+\pi_G(x; z) \leq 0$ . Hence, we obtain orthogonality of  $d^+\pi_G(x; z)$  with  $x - \pi_G(x)$ .  $\square$

LEMMA 21. *Let  $G \subseteq \mathbb{R}^d$  be a non-empty, closed, convex set and  $x \in \mathbb{R}^d$ . Then,  $d^+\pi_G(x; \cdot)$  is positively homogeneous and non-expansive with respect to  $\|\cdot\|$ . Moreover, for all  $z, z' \in \mathbb{R}^d$ ,*

$$(d^+\pi_G(x; z') - d^+\pi_G(x; z))^\top (z' - z) \geq \|d^+\pi_G(x; z') - d^+\pi_G(x; z)\|^2 \geq 0.$$

PROOF. Positive homogeneity is clear from the definition.

Using (13), we have, for all  $z, z' \in \mathbb{R}^d$ ,

$$\begin{aligned} (d^+\pi_G(x; z') - d^+\pi_G(x; z))^\top (z' - z) &= \lim_{\varepsilon \downarrow 0} \varepsilon^{-1} (\pi_G(x + \varepsilon z') - \pi_G(x + \varepsilon z))^\top (z' - z) \\ &\geq \lim_{\varepsilon \downarrow 0} \varepsilon^{-1} \|\pi_G(x + \varepsilon z') - \pi_G(x + \varepsilon z)\|^2 \\ &= \|d^+\pi_G(x; z') - d^+\pi_G(x; z)\|^2. \end{aligned}$$

Finally, non-expansiveness is a direct consequence of the last display, by using Cauchy-Schwarz inequality.  $\square$

LEMMA 22. *Let  $G \subseteq \mathbb{R}^d$  be a non-empty, closed, convex set. Fix  $x \in \mathbb{R}^d$  and let  $f(t) = \|\pi_G(tx)\|$ , for all  $t \geq 0$ . Then,  $f$  is non-decreasing and the map  $t > 0 \mapsto f(t)/t$  is non-increasing.*

In other words, the norm of the projection is non-decreasing and has a non-increasing rate of change along any ray starting at 0. By translating  $G$  (noting that for all  $x_0, x \in \mathbb{R}^d$ ,  $x_0 + \pi_G(x_0 + x) = \pi_{G-x_0}(x)$ ), the lemma also applies to  $f$  of the form  $f(t) = \|x_0 + \pi_G(x_0 + tx)\|$  for any choice of  $x_0, x \in \mathbb{R}^d$ .

PROOF. It is sufficient to show that for all  $x \in \mathbb{R}^d$  and  $t \geq 1$ ,

$$\|\pi_G(x)\| \leq \|\pi_G(tx)\| \leq t\|\pi_G(x)\|.$$

First, Lemma 13 yields the following two sets of inequalities:

$$(15) \quad (x - \pi_G(x))^\top (y - \pi_G(x)), \quad \forall y \in G$$

and

$$(16) \quad (tx - \pi_G(tx))^\top (z - \pi_G(tx)), \quad \forall z \in G.$$

Take  $y = \pi_G(tx)$  and multiply (15) by  $\lambda$ , take  $z = \pi_G(x)$  in (16) and sum the resulting inequalities:

$$(\pi_G(tx) - t\pi_G(x))^\top (\pi_G(tx) - \pi_G(x)) \leq 0.$$

Expanding and using Cauchy-Schwarz inequality imply that

$$\|\pi_G(tx)\|^2 + t\|\pi_G(x)\|^2 - (t+1)\|\pi_G(x)\|\|\pi_G(tx)\| \leq 0.$$

Seeing this inequality as a second degree polynomial inequality in  $\|\pi_G(tx)\|$  yields that

$$\|\pi_G(x)\| \leq \|\pi_G(tx)\| \leq t\|\pi_G(x)\|$$

which is the desired result.  $\square$

Finally, Lemma 22 yields the following property of directional derivatives of projections.

**LEMMA 23.** *Let  $G \subseteq \mathbb{R}^d$  be a non-empty, closed, convex set. Let  $x \in \mathbb{R}^d$  and assume that  $\pi_G$  has directional derivatives at  $x$ . Then,  $\pi_G$  has directional derivatives at every point along the ray from  $\pi_G(x)$  going through  $x$ , that is, at any point of the form  $x_t := \pi_G(x) + t(x - \pi_G(x))$ ,  $t \geq 0$ , and for all  $s, t$  with  $t > s > 0$ , and all  $z \in \mathbb{R}^d$ ,*

$$(17) \quad \|d^+\pi_G(x_t; z)\| \leq \|d^+\pi_G(x_s; z)\|.$$

Note that  $\pi_G$  automatically admits directional derivatives at  $x_0 = \pi_G(x)$ , since  $x_0 \in G$ .

**PROOF.** The existence of directional derivatives at any  $x_t, t > 0$  follows from [43, Proposition 2.2]. Following the proof of that proposition, we also obtain that for all  $t > 0$  and  $z \in \mathbb{R}^d$ ,

$$d^+\pi_G(x_t; z) = d^+\pi_G(x; A_t^{-1}(z))$$

where  $A_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is the bijective map defined as  $A_t = tI_d + (1-t)d^+\pi_G(x; \cdot)$ . In the rest of the proof, let us assume that  $x_0 = \pi_G(x) = 0$ , without loss of generality (we could simply translate  $G$  without affecting the inequality that remains to be proven). Fix  $s, t$  with  $t > s > 0$  and  $z \in \mathbb{R}^d$ . Then, for all  $\varepsilon > 0$ , Lemma 22 yields that

$$\begin{aligned} \frac{\|\pi_G(x_t + \varepsilon z)\|}{\varepsilon} &= \frac{\|\pi_G(tx + \varepsilon z)\|}{\varepsilon} \\ &= \frac{\|\pi_G(t(x + (\varepsilon/t)z))\|}{\varepsilon} \\ &\leq \frac{t}{s} \frac{\|\pi_G(s(x + (\varepsilon/t)z))\|}{\varepsilon} \\ &= \frac{\|\pi_G(x_s + (s\varepsilon/t)z)\|}{s\varepsilon/t} \end{aligned}$$

and taking the limit as  $\varepsilon \rightarrow 0$  implies that  $\|d^+\pi_G(x_t; z)\| \leq \|d^+\pi_G(x_s; z)\|$ .  $\square$

The following result allows to extend (17) to  $s = 0$ .

**LEMMA 24.** *Let  $G \subseteq \mathbb{R}^d$  be a non-empty, closed, convex set. Let  $x \in \mathbb{R}^d$  and assume that  $\pi_G$  has directional derivatives at  $x$ . Then, for all  $z \in \mathbb{R}^d$ ,*

$$\|d^+\pi_G(x; z)\| \leq \|d^+\pi_G(\pi_G(x); z)\|.$$

Note that this lemma is not a consequence of (17) in Lemma 23 because  $s \geq 0 \mapsto \|d^+ \pi_G(x_s; z)\|$ , for fixed  $z \in \mathbb{R}^d$ , is not always continuous at  $s = 0$ . Take, for instance,  $G = B(0, 1)$ ,  $x \in \mathbb{R}^d$  with  $\|x\| > 1$  and  $z = -x$ .

PROOF. Without loss of generality, let us assume that  $0 \in G$  and  $\pi_G(x) = 0$ . First, if  $x \in G$ , then  $\pi_G(x) = x$  and the result is trivial. Assume that  $x \notin G$ . Fix  $z \in \mathbb{R}^d$  and let  $\varepsilon > 0$ . Then, we have

$$\begin{aligned}
 \|\pi_G(x+z)\|^2 &= (x+z)^\top \pi_G(x+z) - (x+z - \pi_G(x+z))^\top \pi_G(x+z) \\
 &\leq (x+z)^\top \pi_G(x+z) \quad \text{by Lemma 13} \\
 &\leq z^\top \pi_G(x+z) \quad \text{by Lemma 13, noting that } \pi_G(x) = 0 \\
 &= z^\top \pi_G(z) + z^\top (\pi_G(x+z) - \pi_G(z)) \\
 &\leq z^\top \pi_G(z) + \pi_G(z)^\top (\pi_G(x+z) - \pi_G(z)) \quad \text{by Lemma 13} \\
 (18) \quad &= (z - \pi_G(z))^\top \pi_G(z) + \pi_G(z)^\top \pi_G(x+z).
 \end{aligned}$$

Now, replacing  $z$  with  $\varepsilon z$  in (18), dividing by  $\varepsilon^2$  and letting  $\varepsilon \downarrow 0$ , we obtain that

$$(19) \quad \|d^+ \pi_G(x; z)\|^2 \leq (z - d^+ \pi_G(0; z))^\top d^+ \pi_G(0; z) + d^+ \pi_G(0; z)^\top d^+ \pi_G(x; z).$$

Note that  $\pi_G$  has directional derivatives at 0 since we have assumed that  $0 \in G$ . Moreover, since  $x \notin G$ ,  $0 = \pi_G(x)$  must be on the boundary of  $G$ . Hence, by Lemma 14,  $d^+ \pi_G(0; \cdot) = \pi_C$  where  $C$  is the support cone to  $G$  at 0. Therefore,  $(z - d^+ \pi_G(0; z))^\top d^+ \pi_G(0; z) = (z - \pi_C(z))^\top \pi_C(z)$ . Now, note that by Lemma 13, for all  $y \in C$ , we have that

$$(z - \pi_C(z))^\top (y - \pi_C(z)) \leq 0.$$

Taking  $y = 0$  on the one hand, and  $y = 2\pi_C(z)$  on the second hand, yields that  $(z - \pi_C(z))^\top \pi_C(z) = 0$ . Therefore, continuing (19), we obtain that

$$\|d^+ \pi_G(x; z)\|^2 \leq d^+ \pi_G(0; z)^\top d^+ \pi_G(x; z)$$

which is bounded by  $\|d^+ \pi_G(0; z)\| \|d^+ \pi_G(x; z)\|$  by Cauchy-Schwarz inequality. The desired result follows readily.  $\square$

## APPENDIX D: ADAPTATION OF THE CONVERGENCE RESULTS FOR $U$ -ESTIMATORS

In this section, we assume that the loss function  $\phi : E^k \times \Theta_0 \rightarrow \mathbb{R}$  is symmetric and measurable in its first  $k$  arguments and convex in its last, and that for all  $\theta \in \Theta_0$ ,  $\phi(\cdot, \theta) \in L^1(P^{\otimes k})$ . This allows to define the population risk  $\Phi(\theta) = \mathbb{E}[\phi(X_1, \dots, X_k, \theta)]$  for all  $\theta \in \Theta_0$ . For all  $n \geq k$ , we define the empirical risk  $\Phi_n$  as  $\Phi_n(\theta) = \frac{1}{\binom{n}{k}} \sum_{1 \leq i_1 < \dots < i_k \leq n} \phi(X_{i_1}, \dots, X_{i_k}, \theta)$ , for all  $\theta \in \Theta_0$ .

For simplicity, for every subset  $I \subseteq \{1, \dots, n\}$  of size  $k$ , we denote by  $X_I$  the vector  $(X_{i_1}, \dots, X_{i_k})$  where  $i_1 < \dots < i_k$  are the elements of  $I$  ordered in increasing order. We also denote by  $\mathcal{P}_{k,n}$  the collection of all subsets of size  $k$  of  $\{1, \dots, n\}$ .

### D.1 Non-differentiable case

Let us prove the analog of Proposition 1 for  $U$ -estimators. Analogs of Theorems 5 and Theorems 6 will follow directly.

PROPOSITION 2. Assume that  $\phi(\cdot, \theta) \in L^2(P^{\otimes k})$  for all  $\theta \in \Theta_0$ . Let  $(\rho_n)_{n \geq 1}$  be any non-decreasing sequence of positive numbers diverging to  $\infty$  as  $n \rightarrow \infty$ . Then, for all  $\theta \in \Theta_0$  and  $t \in \mathbb{R}^d$ ,

$$\rho_n(\Phi_n(\theta + t/\rho_n) - \Phi_n(\theta)) \xrightarrow[n \rightarrow \infty]{} h_{\partial \Phi(\theta)}(t)$$

in probability.

PROOF. Similarly to the proof of Proposition 1, fix  $t \in \mathbb{R}^d$  and define

$$F_n(t) = \rho_n \left( \Phi_n(\theta + t/\rho_n) - \Phi_n(\theta) - \frac{1}{\binom{n}{k} \rho_n} t^\top \sum_{I \in \mathcal{P}_{n,k}} g(X_I, \theta) \right) \\ - \rho_n \left( \Phi(\theta + t/\rho_n) - \Phi(\theta) - \frac{1}{\rho_n} t^\top \mathbb{E}[g(X_1, \dots, X_k, \theta)] \right)$$

and write  $F_n(t) = \sum_{I \in \mathcal{P}_{n,k}} (Z_{I,n} - \mathbb{E}[Z_{I,n}])$  where we set

$$Z_{I,n} = \frac{\rho_n}{\binom{n}{k}} \left( \phi(X_I, \theta + t/\rho_n) - \phi(X_I, \theta) - (1/\rho_n) t^\top g(X_I, \theta) \right)$$

for all  $I \in \mathcal{P}_{n,k}$ . Lemma 11 yields that

$$0 \leq Z_{I,n} \leq \frac{1}{\binom{n}{k}} t^\top (g(X_I, \theta + t/\rho_n) - g(X_I, \theta)) \\ \leq \frac{1}{\binom{n}{k}} t^\top (g(X_I, \theta + t/\rho_1) - g(X_I, \theta))$$

for all  $I \in \mathcal{P}_{n,k}$ . Denoting by  $Y_I = t^\top (g(X_I, \theta + t/\rho_1) - g(X_I, \theta))$  for all  $I \in \mathcal{P}_{n,k}$ , we obtain that for all large enough  $n$  ( $n \geq 2k$  suffices)

$$\text{var}(F_n(t)) = \text{var} \left( \sum_{I \in \mathcal{P}_{n,k}} Z_{I,n} \right) = \sum_{I, J \in \mathcal{P}_{n,k}, I \cap J \neq \emptyset} \text{cov}(Z_{I,n}, Z_{J,n}) \\ \leq \frac{1}{\binom{n}{k}^2} \sum_{I, J \in \mathcal{P}_{n,k}, I \cap J \neq \emptyset} \mathbb{E}[Y_I Y_J] \\ = \frac{1}{\binom{n}{k}^2} \sum_{j=1}^k \binom{n}{k} \binom{k}{j} \binom{n-k}{k-j} \alpha_j \\ = \frac{1}{\binom{n}{k}} \sum_{j=1}^k \binom{k}{j} \binom{n-k}{k-j} \alpha_j$$

where  $\alpha_j = \mathbb{E}[Y_I Y_J]$  for any two sets  $I, J \in \mathcal{P}_{n,k}$  with  $\#(I \cap J) = j$ ,  $j = 1, \dots, k$  ( $\#A$  stands for the cardinality of a set  $A$ ). In the second equality, we used the fact that  $Z_{I,n}$  and  $Z_{J,n}$  are independent if  $I \cap J = \emptyset$ . In the second to last equality, we used the fact that for  $j = 1, \dots, k$ , the number of pairs of sets  $I, J \in \mathcal{P}_{n,k}$  with  $\#(I \cap J) = j$  is  $\binom{n}{k} \binom{k}{j} \binom{n-k}{k-j}$  (choose  $I$  first, then  $j$  elements in  $I$  and  $k-j$  outside of  $I$  to obtain  $J$ ).

Note that  $\alpha_1, \dots, \alpha_k$  do not depend on  $n$ , and each term in the product is of order at most  $1/n$ . Hence,  $\text{var}(F_n(t)) \xrightarrow{n \rightarrow \infty} 0$  so  $F_n(t) \xrightarrow{n \rightarrow \infty} 0$  in probability.

By Theorem 8,  $\frac{1}{\binom{n}{k}} \sum_{I \in \mathcal{P}_{n,k}} g(X_I, \theta) \xrightarrow{n \rightarrow \infty} \mathbb{E}[g(X_1, \dots, X_k, \theta)]$  almost surely, hence, in probability, and Lemma 9 yields that  $\rho_n (\Phi(\theta + t/\rho_n) - \Phi(\theta)) \xrightarrow{n \rightarrow \infty} h_{\partial\Phi(\theta)}(t)$ . Hence, we obtain the desired result.  $\square$

## D.2 Proof of Theorem D

As in the proof of Theorem 7, fix  $R > 0$  and let

$$F_n(t) = n \left( \Phi_n(\theta^* + t/\sqrt{n}) - \Phi_n(\theta^*) \right) - \left( \frac{\sqrt{n}}{\binom{n}{k}} t^\top \sum_{I \in \mathcal{P}_{n,k}} g(X_I, \theta^*) + \frac{1}{2} t^\top S t \right)$$

for all  $t \in B_S(0, R)$ , for all large enough  $n$  so  $B_S(\theta^*, R/\sqrt{n}) \subseteq \Theta_0$ , and where  $S = \nabla^2 \Phi(\theta^*)$ . Let us show that for all  $t \in B_S(0, R)$ ,  $F_n(t) \xrightarrow[n \rightarrow \infty]{} 0$  in probability. For this, we let

$$Z_{I,n} = \frac{n}{\binom{n}{k}} \left( \phi(X_I, \theta^* + t/\sqrt{n}) - \phi(X_I, \theta^*) - \frac{t^\top}{\sqrt{n}} g(X_I, \theta^*) \right)$$

for each  $I \in \mathcal{P}_{n,k}$ . Now, we note that for each  $I \in \mathcal{P}_{n,k}$ ,

$$0 \leq Z_{I,n} \leq \frac{\sqrt{n}}{\binom{n}{k}} \left( g(X_I, \theta^* + t/\sqrt{n}) - g(X_I, \theta^*) \right)$$

thanks to Lemma 11. Setting  $Y_{I,n} = g(X_I, \theta^* + t/\sqrt{n}) - g(X_I, \theta^*)$  for all  $I \in \mathcal{P}_{n,k}$ , we obtain:

$$\begin{aligned} \text{var} \left( \sum_{I \in \mathcal{P}_{n,k}} Z_{I,n} \right) &= \sum_{I, J \in \mathcal{P}_{n,k}, I \cap J \neq \emptyset} \text{cov}(Z_{I,n}, Z_{J,n}) \\ &\leq \frac{n}{\binom{n}{k}^2} \sum_{I, J \in \mathcal{P}_{n,k}, I \cap J \neq \emptyset} \mathbb{E}[Y_{I,n} Y_{J,n}] \\ &= \frac{n}{\binom{n}{k}} \sum_{j=1}^k \binom{k}{j} \binom{n-k}{k-j} a_{j,n} \end{aligned}$$

where, for all  $j = 1, \dots, k$ ,  $a_{j,n} = \mathbb{E}[Y_{I,n} Y_{J,n}]$  for any fixed  $I, J \in \mathcal{P}_{n,k}$  with  $\#(I \cap J) = j$ . Fix  $I_0 = \{1, \dots, k\}$  and  $J_0 = \{1, \dots, j, k+1, \dots, 2k-j\}$ . Now, just as in the proof of Theorem 7, note that  $(Y_{I_0,n})_{n \geq 1}$  is a non-increasing sequence (by Lemma 11) of non-negative random variables, hence, it converges almost surely to some non-negative random variable  $Y_{I_0}$ . By monotone convergence, we must then have that  $\mathbb{E}[Y_{I_0,n}] \xrightarrow[n \rightarrow \infty]{} \mathbb{E}[Y_{I_0}]$ . Yet,  $\mathbb{E}[Y_{I_0,n}] = \nabla \Phi(\theta^* + t/\sqrt{n}) - \nabla \Phi(\theta^*)$ , which goes to 0 as  $n \rightarrow \infty$ , since  $\Phi$  is twice differentiable at  $\theta^*$ , hence  $\nabla \Phi$  is continuous at  $\theta^*$ . Therefore,  $\mathbb{E}[Y_{I_0}] = 0$ , which yields that  $Y_{I_0} = 0$  almost surely. Similarly,  $Y_{J_0,n} \xrightarrow[n \rightarrow \infty]{} 0$  almost surely and, now, monotone convergence implies that  $\mathbb{E}[Y_{I_0,n} Y_{J_0,n}] \xrightarrow[n \rightarrow \infty]{} 0$ . Finally, we obtain that each  $a_{j,n} \xrightarrow[n \rightarrow \infty]{} 0$ ,  $j = 1, \dots, k$ . Moreover, for each  $j = 1, \dots, k$ ,  $\binom{k}{j} \binom{n-k}{k-j}$  is of the same order as  $n^j$  as  $n \rightarrow \infty$  so we readily obtain that

$$\text{var} \left( \sum_{I \in \mathcal{P}_{n,k}} Z_{I,n} \right) \xrightarrow[n \rightarrow \infty]{} 0.$$

The fact that  $F_n(t) \xrightarrow[n \rightarrow \infty]{} 0$  in probability then follows from  $\phi$  being twice differentiable at  $\theta^*$ .

Now, the rest of the proof is almost identical to that of Theorem 7, with the only difference that, using Theorem 9, an extra  $k$  factor will appear in the asymptotic behavior of the minimizer  $t_n^*$  of

$$\frac{\sqrt{n}}{\binom{n}{k}} t^\top \sum_{I \in \mathcal{P}_{n,k}} g(X_I, \theta^*) + \frac{1}{2} t^\top S t.$$

## APPENDIX E: MISCELLANEOUS RESULTS

Let us give yet a second corollary to Lemma 2, that allows to go from pointwise to uniform convergence, in  $L^p$  sense ( $p \geq 1$ ).

**COROLLARY 2.** *Let  $p \geq 1$ . Let  $f, f_1, f_2, \dots$  be random convex functions defined on an open convex set  $G_0 \subseteq \mathbb{R}^d$ . Assume that for all  $n \geq 1$  and all  $t \in G_0$ ,  $f_n(t) \in L^p(\mathbb{P})$ . Assume also that  $\mathbb{E}[|f_n(t) - f(t)|^p] \xrightarrow{n \rightarrow \infty} 0$  for all  $t \in G_0$ . Then, for all compact sets  $K \subseteq G_0$ ,  $\mathbb{E}[\sup_K |f_n - f|^p] \xrightarrow{n \rightarrow \infty} 0$ .*

**PROOF.** Let  $K \subseteq G_0$  be a compact set. Since  $K$  is compact, so is its convex hull, by [49, Theorem 1.1.11]. Hence, without loss of generality, in the sequel, let us assume that  $K$  is convex.

Since  $G_0$  is open, there exists  $\eta > 0$  satisfying that  $K^{2\eta} := \{x \in \mathbb{R}^d : d(x, K) \leq 2\eta\} \subseteq G_0$ . Moreover, there exists a convex polytope  $P$  with  $K^\eta \subseteq P \subseteq K^{2\eta}$ , see [10]. Let  $v_1, \dots, v_r$  ( $r \geq 1$ ) the vertices of  $P$ .

Fix  $\varepsilon > 0$  with  $\varepsilon \leq \eta/2$  and let  $t_1, \dots, t_N \in K$  (with  $N \geq 1$ ) be an  $\varepsilon$ -approximation of  $K$ , that is, such that for all  $t \in K$ ,  $\|t - t_j\| \leq \varepsilon$  for some  $j \in \{1, \dots, N\}$ .

Let  $t \in K$  and  $j \in \{1, \dots, N\}$  satisfying  $\|t - t_j\| \leq \varepsilon$ . Assume for now that  $t \neq t_j$  and let  $z_-$  and  $z_+$  be the two points at the intersection of  $\partial P$  and the line passing through  $t$  and  $t_j$ , that is,

$$z_- = t_j + \lambda_-(t - t_j)$$

and

$$z_+ = t + \lambda_+(t_j - t)$$

for some  $\lambda_-, \lambda_+ \geq 1$ . Note that  $\lambda_- = \frac{\|z_- - t_j\|}{\|t - t_j\|} = \frac{\|z_- - t\|}{\|t - t_j\|} + 1 \geq 1 + \frac{\eta}{\varepsilon} > 1$  and that  $\lambda_+ = \frac{\|z_+ - t\|}{\|t - t_j\|} = \frac{\|z_+ - t_j\|}{\|t - t_j\|} + 1 \geq 1 + \frac{\eta}{\varepsilon} > 1$ . For each  $n \geq 1$ , convexity of  $f_n$  yields, on the one hand, that

$$(20) \quad f_n(t) - f_n(t_j) \leq (1/\lambda_-)(f_n(z_-) - f_n(t_j)) \leq (1/\lambda_-) \left( \max_P f_n - f_n(t_j) \right) \leq \frac{\varepsilon}{\eta} \left( \max_P f_n - f_n(t_j) \right)$$

and that

$$f_n(t_j) \leq (1 - 1/\lambda_+)f_n(t) + (1/\lambda_+)f_n(z_+),$$

so, dividing both sides by  $(1 - 1/\lambda_+)$  and subtracting  $f_n(t_j)$ ,

$$\begin{aligned} f_n(t) - f_n(t_j) &\geq (1/\lambda_+)(1 - 1/\lambda_+)^{-1} (f_n(t_j) - f_n(z_+)) \\ &= \frac{1}{\lambda_+ - 1} (f_n(t_j) - f_n(z_+)) \\ &\geq \frac{1}{\lambda_+ - 1} \left( f_n(t_j) - \max_P f_n \right) \\ (21) \quad &\geq \frac{\varepsilon}{\eta} \left( f_n(t_j) - \max_P f_n \right). \end{aligned}$$

Let  $M_n = \max(f_n(v_1), \dots, f_n(v_r))$  and  $m_n = \min(f_n(t_1), \dots, f_n(t_N))$ . Convexity of  $f_n$  yields that  $\max_P f_n \leq M_n$  and we obtain, from (20) and (21), that

$$|f_n(t) - f_n(t_j)| \leq \frac{\varepsilon}{\eta} (M_n - m_n).$$

Similarly for  $f$ , we have that

$$|f(t) - f(t_j)| \leq \frac{\varepsilon}{\eta} (M - m)$$



where we set  $M = \max(f(v_1), \dots, f(v_p))$  and  $m = \min(f(t_1), \dots, f(t_N))$ .

Finally, writing  $|f_n(t) - f(t)| \leq |f_n(t) - f_n(t_j)| + |f_n(t_j) - f(t_j)| + |f(t) - f(t_j)|$ , we obtain:

$$\sup_{t \in K} |f_n(t) - f(t)| \leq \frac{\varepsilon}{\eta} (M_n - m_n + M - m) + \max_{1 \leq j \leq N} |f_n(t_j) - f(t_j)|.$$

Now, raising to the power  $p$  and taking the expectation on both sides, we obtain that

$$\begin{aligned} \mathbb{E}[\sup_{t \in K} |f_n(t) - f(t)|^p] &\leq \frac{5^{p-1}\varepsilon}{\eta} (\mathbb{E}[|M_n|^p] + \mathbb{E}[|m_n|^p] + \mathbb{E}[|M|^p] + \mathbb{E}[|m|^p]) \\ &\quad + 5^{p-1} \sum_{j=1}^N \mathbb{E}[|f_n(t_j) - f(t_j)|^p] \end{aligned}$$

where we used the fact that  $(a_1 + a_2 + a_3 + a_4 + a_5)^p \leq 5^{p-1}(a_1^p + a_2^p + a_3^p + a_4^p + a_5^p)$  for all positive numbers,  $a_1, a_2, a_3, a_4, a_5$  and we bounded the maximum of non-negative numbers by their sum in the last term. Now, the assumption implies that, for large enough, each term in the last sum can be bounded by  $\varepsilon/N$  and, hence, the whole right hand side can be bounded by  $C\varepsilon$  for some positive constant  $C$ . Since  $\varepsilon > 0$  was any (small enough) positive number, this implies that  $\mathbb{E}[\sup_{t \in K} |f_n(t) - f(t)|^p] \xrightarrow[n \rightarrow \infty]{} 0$ .  $\square$

LEMMA 25. *Let  $(Z_n)_{n \geq 1}$  be a sequence of real random variables satisfying that  $\rho_n Z_n \xrightarrow[n \rightarrow \infty]{} 0$  in probability, for any choice of non-decreasing sequence  $(\rho_n)_{n \geq 1}$  of positive numbers, diverging to  $\infty$  as  $n \rightarrow \infty$ . Then,  $P(Z_n \neq 0) \xrightarrow[n \rightarrow \infty]{} 0$ .*

PROOF. Assume, for the sake of contradiction, that  $P(Z_n \neq 0)$  does not go to 0 as  $n \rightarrow \infty$ . That is, there is some  $\varepsilon$  and an increasing sequence  $(k_n)_{n \geq 1}$  of positive integers such that  $P(Z_{k_n} \neq 0) \geq \varepsilon$  for all  $n \geq 1$ . For each  $n \geq 1$ , since the map  $t \in \mathbb{R} \mapsto P(|Z_{k_n}| > t)$  is right-continuous, so there must exist some  $\alpha_n > 0$  with the property that  $P(|Z_{k_n}| > \alpha_n) \geq \varepsilon/2$ . Since  $Z_n \xrightarrow[n \rightarrow \infty]{} 0$  in probability by assumption, the sequence  $(\alpha_n)_{n \geq 1}$  must converge to 0. Since we could extract a non-increasing subsequence of it, let us assume, for simplicity, that  $(\alpha_n)_{n \geq 1}$  is non-increasing. Then, one can choose a non-decreasing sequence  $(\rho_n)_{n \geq 1}$  of positive numbers, such that  $\rho_{k_n} = 1/\alpha_n$  for all  $n \geq 1$ . The assumption implies that  $\rho_{k_n} Z_{k_n}$  must converge to 0 in probability, as a subsequence of  $(\rho_n Z_n)_{n \geq 1}$ . Since this is not the case by construction, we obtain a contradiction.  $\square$