Calhoun: The NPS Institutional Archive

DSpace Repository

Faculty and Researchers | Faculty and Researchers' Publications

2020-02

Set-Convergence and Its Application: A Tutorial

Royset, Johannes O.

ArXiv

Royset, Johannes O. "Set-Convergence and Its Application: A Tutorial." arXiv preprint arXiv:2002.09774 (2020).
http://hdl.handle.net/10945/65314

This publication is a work of the U.S. Government as defined in Title 17, United States Code, Section 101. Copyright protection is not available for this work in the United States.

Downloaded from NPS Archive: Calhoun

Calhoun is the Naval Postgraduate School's public access digital repository for research materials and institutional publications created by the NPS community. Calhoun is named for Professor of Mathematics Guy K. Calhoun, NPS's first appointed -- and published -- scholarly author.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

http://www.nps.edu/library

arXiv:2002.09774v1 [math.OC] 22 Feb 2020

Set-Convergence and Its Application: A Tutorial

Johannes O. Royset
Operations Research Department
Naval Postgraduate School, Monterey, California

Abstract. Optimization problems, generalized equations, and the multitude of other variational problems invariably lead to the analysis of sets and set-valued mappings as well as their approximations. We review the central concept of set-convergence and explain its role in defining a notion of proximity between sets, especially for epigraphs of functions and graphs of set-valued mappings. The development leads to an approximation theory for optimization problems and generalized equations with profound consequences for the construction of algorithms. We also introduce the role of set-convergence in variational geometry and subdifferentiability with applications to optimality conditions. Examples illustrate the importance of set-convergence in stability analysis, error analysis, construction of algorithms, statistical estimation, and probability theory.

Keywords: set-convergence, epi-convergence, graphical convergence, weak convergence, stability, approximation theory, variational geometry, subdifferentiability, truncated Hausdorff distance.

Date: February 25, 2020

1 Introduction

In the study of optimization problems, we quickly encounter a plethora of sets that might cause concerns for the classically trained: Sets of solutions are rarely singletons, local properties aren't characterized by a single gradient vector but sets of subgradients, and the accuracy of function approximations is determined by the distance between certain sets. More general problems such as variational inequalities and generalized equations are even defined in terms of set-valued mappings. The mathematics of sets is therefore central to most analysis of optimization and variational problems beyond the elementary cases. In particular, a notion of convergence of sets becomes necessary, for example, to clarify what it means for a sequence of solution sets or a sequence of subgradient sets to converge as well as for many other concepts.

In 1902, Painlevé defined set-convergence in the sense we have it today, with Hausdorff and Kuratowski supporting its dissemination by including the topic in their books in 1927 and 1933, respectively[1]. A set can be viewed as a point in a space of sets on which a topology and other mathematical structure can be defined. For the majority of the 20th century, set-convergence was studied from this point of view culminating with Beer's monograph [7]. The growing number of variational problems in statistics, medicine, economics, business, engineering, and the sciences now drives a renewed interest in set-convergence from a more applied angle. In this article, we summarize key properties of set-convergence and highlight their importance when analyzing optimization problems and generalized equations as well as corresponding algorithms.

[1]Detailed historical remarks can be found at the end of Chapter 4 in [18].

An illustrative example arises in the classical penalty method for constrained optimization. For $f_0, h : \mathbb{R}^n \to \mathbb{R}$, the problem of minimizing $f_0(x)$ subject to $h(x) = 0$ can be stated concisely as[2]

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, f(x) = f_0(x) + \iota_{\{0\}}\big(h(x)\big).$$

Since an infinite penalty isn't compatible with direct application of standard algorithms, the penalty method considers the alternative problems

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, f^\nu(x) = f_0(x) + \theta^\nu\big(h(x)\big)^2, \quad \text{with } \theta^\nu > 0, \quad \nu \in \mathbb{N} = \{1, 2, \dots\}.$$

At least for smooth f_0 and h, standard unconstrained algorithms apply to the alternative problems but can we guarantee that their minimizers are near those of the actual problem? Intuitively, this relates to whether f^ν is near f in some sense. A classical treatment is complicated by the fact that $f(x) - f^\nu(x) = \infty$ whenever $h(x) \neq 0$ regardless of the value of θ^ν. Thus, it appears that the difference between f and f^ν remains large even as $\theta^\nu \to \infty$ and one might be led to believe that the alternative problems aren't suitable approximations of the actual problem.

For another example, consider the problem of minimizing $f_0(x) = -x$ subject to $g(x) = x^3 - x^2 - x + 1 \leq 0$; the feasible set is $(-\infty, -1] \cup \{1\}$ and $x^\star = 1$ is the unique minimizer. If g is replaced by the seemingly accurate approximation $g^\nu = g + \nu^{-1}$ so that $\sup_{x\in\mathbb{R}} |g^\nu(x) - g(x)| = \nu^{-1}$, then one might expect minimizers to change just a little. However, we obtain a substantially different minimizer near -1 even for large ν. In these two examples, an assessment based on differences in function values is misleading. The key insight is provided by the epigraphs of the alternative functions and their proximity to that of the actual function.

For a function $f : \mathbb{R}^n \to \overline{\mathbb{R}} = [-\infty, \infty]$, we recall that the *epigraph* $\operatorname{epi} f = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq \alpha\}$. Moreover, the *minimum value* of f is $\inf f = \inf\{f(x) \mid x \in \mathbb{R}^n\}$ and the *set of minimizers* is $\operatorname{argmin} f = \{x \in \operatorname{dom} f \mid f(x) = \inf f\}$, where $\operatorname{dom} f = \{x \in \mathbb{R}^n \mid f(x) < \infty\}$.

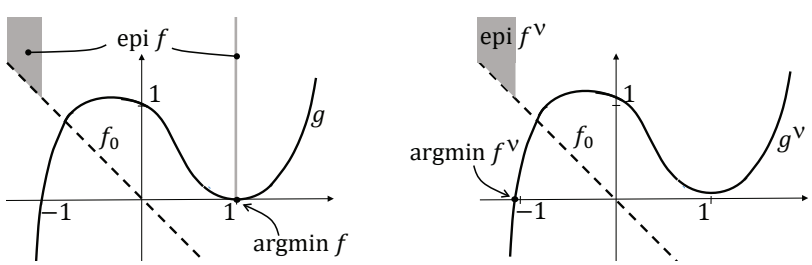

Figure 1: Constraint function $g(x) = x^3 - x^2 - x + 1$ is approached uniformly by $g^\nu = g + \nu^{-1}$, but minimum value and minimizer of $f^\nu = f_0 + \iota_{(-\infty,0]}(g^\nu(\cdot))$ aren't close to those of $f = f_0 + \iota_{(-\infty,0]}(g(\cdot))$.

In the second example above, we compare the actual problem of minimizing $f = f_0 + \iota_{(-\infty,0]}(g(\cdot))$ with that of minimizing $f^\nu = f_0 + \iota_{(-\infty,0]}(g^\nu(\cdot))$. Figure 1 shows $\operatorname{epi} f$ and $\operatorname{epi} f^\nu$. The two epigraphs are rather different and they don't get close in any reasonable sense; $\operatorname{epi} f$ contains a vertical ray starting

[2]For any set C, $\iota_C(u) = 0$ if $u \in C$ and $\iota_C(u) = \infty$ otherwise.

at $(1,0)$ that is distinct from any points in $\operatorname{epi} f^\nu$. The effect is that $\operatorname{argmin} f^\nu$ and $\inf f^\nu$ deviate greatly from $\operatorname{argmin} f$ and $\inf f$, respectively. Indeed, the proximity between epigraphs provides a robust assessment of whether the corresponding solutions are close as we see below.

To formalize set-convergence, we define the *point-to-set distance* between $\bar{x} \in \mathbb{R}^n$ and $C \subset \mathbb{R}^n$ as

$$\operatorname{dist}_2(\bar{x}, C) = \inf_{x \in C} \|x - \bar{x}\|_2 \ \text{ when } C \neq \emptyset \ \text{ and } \ \operatorname{dist}_2(\bar{x}, \emptyset) = \infty. \tag{1.1}$$

1.1 Definition (set-convergence[3]). *For $\{C, C^\nu \subset \mathbb{R}^n,\ \nu \in \mathbb{N}\}$, we say that C^ν set-converges to C, written $C^\nu \xrightarrow{s} C$ or $\operatorname{Lim} C^\nu = C$, when*

$$C \text{ is closed and } \operatorname{dist}_2(x, C^\nu) \to \operatorname{dist}_2(x, C) \text{ for all } x \in \mathbb{R}^n.$$

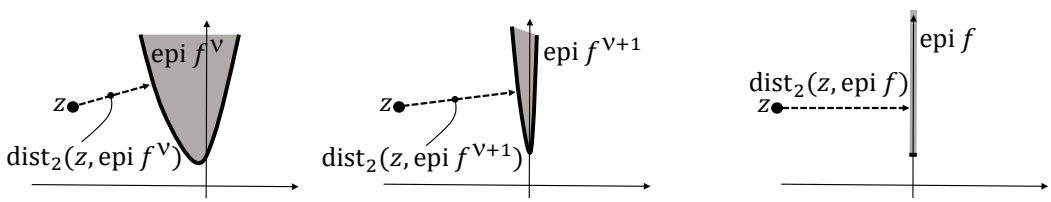

Figure 2: The epigraphs of f^ν set-converge to $\operatorname{epi} f = \{0\} \times [1, \infty)$.

Figure 2 visualizes the definition applied to $\operatorname{epi} f^\nu$ and $\operatorname{epi} f$ in the first example above when $f_0(x) = (x+1)^2$ and $h(x) = x$. Indeed, $\operatorname{dist}_2(z, \operatorname{epi} f^\nu) \to \operatorname{dist}_2(z, \operatorname{epi} f)$ for all $z \in \mathbb{R}^{n+1}$. As expected from the penalty method, we see that $\operatorname{argmin} f^\nu$, which is a bit to the left of the origin, tends to $\operatorname{argmin} f = \{0\}$.

In general, for any $f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a proper[4] function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$:

$$\operatorname{epi} f^\nu \xrightarrow{s} \operatorname{epi} f \text{ and } \bar{x} \text{ is a cluster point of } \{x^\nu \in \operatorname{argmin} f^\nu, \nu \in \mathbb{N}\} \implies \bar{x} \in \operatorname{argmin} f.$$

This fact, further elaborated on in §3, guides the construction of algorithms such as the penalty method and helps us reformulate problems that are fundamentally unstable. As an example, consider

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ f_0(x) \text{ subject to } g_i(x) \leq 0,\ i = 1, \ldots, q,$$

where the individual functions need to be replaced by approximations for computational reasons or because they are only partially known. Regardless, let $f_0^\nu, g_i^\nu, \ldots, g_q^\nu$ be the approximations and, for $\alpha^\nu \geq 0$, suppose that $\sup_{x \in \mathbb{R}^n} |f_0^\nu(x) - f_0(x)| \leq \alpha^\nu$ and $\sup_{x \in \mathbb{R}^n} \max_{i=1,\ldots,q} |g_i^\nu(x) - g_i(x)| \leq \alpha^\nu$, with all the functions being real-valued. As discussed in conjuncture with Figure 1, one can't expect that substituting in f_0^ν for f_0 and g_i^ν for g_i produce a problem with minimizers near those of the actual one even if α^ν is small. The goal becomes to construct alternative functions with epigraphs that set-converge to one corresponding to the actual problem as $\alpha^\nu \to 0$ because then minimizers will follow suit.

[3]This notion is also referred to as Painlevé-Kuratowski convergence.

[4]A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *proper* if $f(x) > -\infty$ for all $x \in \mathbb{R}^n$ and $f(x) < \infty$ for some $x \in \mathbb{R}^n$.

An alternative and indeed approximating problem emerges by softening the constraints and using a penalty $\theta^\nu > 0$:

$$\underset{x\in\mathbb{R}^n, y\in\mathbb{R}^q}{\text{minimize}} \;\; f_0^\nu(x) + \theta^\nu \sum\nolimits_{i=1}^q y_i \;\; \text{subject to} \;\; g_i^\nu(x) \le y_i, \;\; 0 \le y_i, \;\; i = 1, \ldots, q.$$

Although seemingly rather different than the actual problem, it's epigraphically close in the sense we seek. Specifically, the approximating problem can be written compactly as minimizing the function $f^\nu : \mathbb{R}^n \times \mathbb{R}^q \to \overline{\mathbb{R}}$ given by

$$f^\nu(x,y) = f_0^\nu(x) + \sum\nolimits_{i=1}^q \varphi^\nu(y_i) + \sum\nolimits_{i=1}^q \iota_{(-\infty,0]}\big(g_i^\nu(x) - y_i\big), \text{ where } \varphi^\nu(\beta) = \begin{cases} \theta^\nu \beta & \text{if } \beta \ge 0 \\ \infty & \text{otherwise.} \end{cases}$$

The actual problem is expressed in terms of x only, but to allow comparison with the approximation we artificially introduce y and state it as minimizing the function $f : \mathbb{R}^n \times \mathbb{R}^q \to \overline{\mathbb{R}}$ defined by

$$f(x,y) = f_0(x) + \sum\nolimits_{i=1}^q \iota_{\{0\}}(y_i) + \sum\nolimits_{i=1}^q \iota_{(-\infty,0]}\big(g_i(x) - y_i\big).$$

Trivially, y must be the zero vector for f to be finite so this is indeed an equivalent statement of the actual problem. The similarities between f^ν and f are clear. In particular, $\operatorname{epi}\varphi^\nu \xrightarrow{s} \operatorname{epi}\iota_{\{0\}}$ when $\theta^\nu \to \infty$. Using the techniques of §3, we can show that $\operatorname{epi} f^\nu \xrightarrow{s} \operatorname{epi} f$ under the assumptions that f_0 is continuous, g_i is lower semicontinuous (lsc) for all i, $\theta^\nu \to \infty$, and $\theta^\nu \alpha^\nu \to 0$. Practitioners often soften constraints and the above analysis reveals that there are good reasons for this: the resulting problem becomes a valid approximation of an actual problem, which may not even be fully known. In contrast, the naive "approximation" obtained by simply substituting in f_0^ν and g_i^ν can be arbitrarily poor.

These introductory examples illustrate the role of set-convergence in analysis of optimization problems and the construction of numerical procedures for their solution. After some background on set-convergence in §2, we return to that subject in §3. Set-convergence also enters in the development of robust notions of normal cones and subgradients that reach beyond the convex case as seen in §4. Continuity and approximation of set-valued mappings are intimately tied to set-convergence, with significant implications for the solution of generalized equations as discussed in §5. Quantification of set-convergence enables us to obtain rates of convergence for algorithms and bound the solution error due to approximations in optimization problems and generalized equations; see §6. Applications of set-convergence in probability and statistics are recorded in §3. The article ends in §7 with a summary of topics not covered including extensions to infinite dimensions.

2 Inner and Outer Limits

Set-convergence can be understood and fully characterized using the notions of outer and inner limits, which in fact was the approach of Painlevé. These are not only useful stepping stones towards set-convergence, but also furnish intermediate properties that might suffice in a particular application.

The *inner limit* of $\{C^\nu \subset \mathbb{R}^n, \nu \in \mathbb{N}\}$, denoted by $\operatorname{LimInn} C^\nu$, is the collection of limit points to which sequences of points selected from the sets converge. Specifically,

$$\operatorname{LimInn} C^\nu = \big\{x \in \mathbb{R}^n \;\big|\; \exists x^\nu \in C^\nu \to x\big\}.$$

For example, the inner limit of the line segments $C^\nu = [1/\nu, 2/\nu]$ is $\{0\}$. A slight clarification regarding empty sets is in place. Suppose that $C^1 = \emptyset$, but $C^\nu = [1/\nu, 2/\nu]$ for $\nu \geq 2$. Still, $\operatorname{LimInn} C^\nu = \{0\}$. Since convergence of a sequence is always determined by its tail and not a finite number of its elements, we aren't concerned about not being able to select a point from C^1. It suffices in the definition of inner limits to have points in C^ν converging to x for all ν sufficiently large.

Figure 3 furnishes two more examples: On the left, $\operatorname{LimInn} C^\nu$ is the third quadrant. On the right, collections of vertical line segments C^ν eventually "fill" a rectangle.

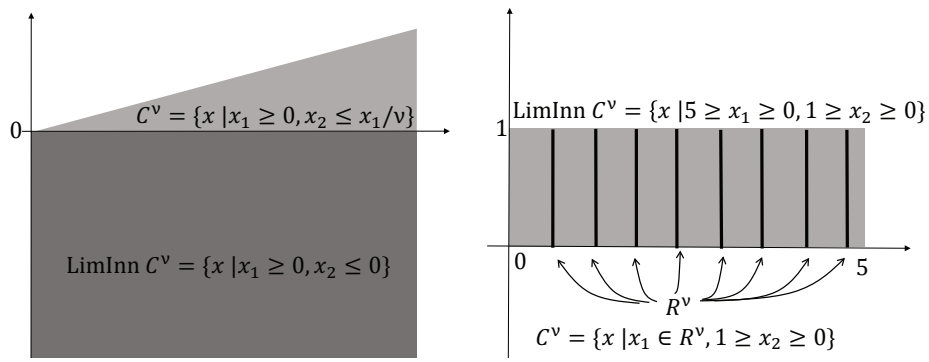

Figure 3: Examples of inner limits; R^ν is the first ν numbers in an ordered list of the rational numbers in $[0, 5]$.

If $C^\nu = \{0\}$ when ν is odd and $C^\nu = [0, 1]$ when ν is even, then $\operatorname{LimInn} C^\nu = \{0\}$. The inner limit isn't $[0, 1]$ because we can't construct $x^\nu \in C^\nu$ converging to $x \in (0, 1]$. This brings us to outer limits.

The *outer limit* of $\{C^\nu \subset \mathbb{R}^n, \nu \in \mathbb{N}\}$, denoted by $\operatorname{LimOut} C^\nu$, is the collection of cluster points to which subsequences of points selected from the sets converge. To write this formally, we use the following notation for subsequences:

$$\mathcal{N}_\infty^\# \text{ is the set of all infinite collections of increasing numbers from } \mathbb{N}.$$

For example, $\{1, 3, 5, \dots\} \in \mathcal{N}_\infty^\#$ and $\{2, 4, 6, \dots\} \in \mathcal{N}_\infty^\#$. A subsequence of $\{x^\nu \in \mathbb{R}^n, \nu \in \mathbb{N}\}$ is then of the form $\{x^\nu \in \mathbb{R}^n, \nu \in N\}$ for some $N \in \mathcal{N}_\infty^\#$, with $x^\nu \xrightarrow[N]{} x$ expressing that x is a cluster point corresponding to the subsequence specified by the index set N. In this notation,

$$\operatorname{LimOut} C^\nu = \left\{x \in \mathbb{R}^n \;\middle|\; \exists N \in \mathcal{N}_\infty^\# \text{ and } x^\nu \in C^\nu \xrightarrow[N]{} x\right\}.$$

Hence, $x \in \operatorname{LimOut} C^\nu$ if we can select an index set $N \in \mathcal{N}_\infty^\#$ and points $\{x^\nu \in C^\nu, \nu \in N\}$ converging to x. In the example $C^\nu = \{0\}$ when ν is odd and $C^\nu = [0, 1]$ when ν is even, $\operatorname{LimOut} C^\nu = [0, 1]$ because for any $x \in [0, 1]$, we can rely on $N = \{2, 4, \dots\}$ and select $x^\nu = x$ for $\nu \in N$. The relation between the inner and the outer limits in this example points to a fact that motivates the terminology "inner" and "outer:" Generally, $\operatorname{LimInn} C^\nu \subset \operatorname{LimOut} C^\nu$. Although $\operatorname{Lim} C^\nu$ may not exist (as is clear from the "odd-even" example above), the inner limit and the outer limit always exist but could be the empty set. Moreover, they are closed sets and characterize set-convergence [18, Section 4.B]:

2.1 Proposition (inner and outer limits). *For* $\{C, C^\nu \subset \mathbb{R}^n, \nu \in \mathbb{N}\}$,

$$\operatorname{Lim} C^\nu = C \quad \Longleftrightarrow \quad \operatorname{LimInn} C^\nu = \operatorname{LimOut} C^\nu = C.$$

In applications, a sequence of sets may not set-converge but useful insight can come from the inner and/or outer limits. For example, given a closed feasible set $C \subset \mathbb{R}^n$ that is approximated by $\{C^\nu \subset \mathbb{R}^n, \nu \in \mathbb{N}\}$, suppose that an algorithm generates $\{x^\nu \in C^\nu, \nu \in \mathbb{N}\}$, which we hope are near a feasible point in C. Indeed, if $\operatorname{LimOut} C^\nu \subset C$, then every cluster point of $\{x^\nu, \nu \in \mathbb{N}\}$ is contained in C and is thus feasible for the actual problem. This might suffice as justification for the approximating sets C^ν. If $C^\nu \xrightarrow{s} C$, then we have in addition that every $x \in C$ can be approached by points in C^ν, i.e., there is $x^\nu \in C^\nu \to x$.

Practical verification of set-convergence is supported by a collection of rules that enable us to conclude that images, inverse images, intersections, unions, sums, products, and other operations on sets preserve set-convergence, at least in part, under various assumptions. We refer to [18, Section 4.G] for a detailed treatment and here concentrate on the central cases of intersections and unions.

A feasible set in an optimization problem is often specified as the intersection and/or union of a collection of other sets, each representing some requirement, which may need approximation. The question then becomes: how do approximations of the "component" sets affect the feasible set?

2.2 Proposition (union and intersection under set-convergence). *For sets* $\{C^\nu, D^\nu \subset \mathbb{R}^n, \nu \in \mathbb{N}\}$,

$$\operatorname{Lim}\left(C^\nu \cup D^\nu\right) \xrightarrow{s} \operatorname{Lim} C^\nu \cup \operatorname{Lim} D^\nu$$
$$\operatorname{LimOut}\left(C^\nu \cap D^\nu\right) \subset \operatorname{LimOut} C^\nu \cap \operatorname{LimOut} D^\nu.$$

If all C^ν, D^ν *are convex and*[5] $\operatorname{int}(\operatorname{Lim} C^\nu) \cap \operatorname{Lim} D^\nu \neq \emptyset$, *then* $\operatorname{Lim}(C^\nu \cap D^\nu) \xrightarrow{s} \operatorname{Lim} C^\nu \cap \operatorname{Lim} D^\nu$.

Proof. The first two statements follow readily from the definitions of inner limits and outer limits. For the convex case, consult [18, Theorem 4.32]. □

The situation is less than ideal for intersections: $\operatorname{LimOut}(C^\nu \cap D^\nu)$ can be strictly contained in $\operatorname{Lim} C^\nu \cap \operatorname{Lim} D^\nu$ even for convex sets. For example, $C^\nu = [-1, -1/\nu]$ and $D^\nu = [1/\nu, 1]$ has $\operatorname{LimOut}(C^\nu \cap D^\nu) = \emptyset$, while $\operatorname{LimOut} C^\nu \cap \operatorname{LimOut} D^\nu = \{0\}$. This highlights the challenge in constrained optimization already indicated in the Introduction: A tiny change of the components of a problem may have large effects on the feasible set. Set-convergence helps us to identify trouble. If a feasible set $C \cap D$ isn't stable under perturbations of C and D in the sense that approximations $C^\nu \cap D^\nu$ fail to set-converge to $C \cap D$ despite $C^\nu \xrightarrow{s} C$ and $D^\nu \xrightarrow{s} D$, then optimization over $C \cap D$ is somehow ill-posed. Any small misspecification of C or D can cause misleading solutions. In the convex case, we achieve such stability if C and D overlap "sufficiently" as stated in the proposition.

Unions of sets are well-behaved in this sense, which has the following ramification for min-functions.

[5] We denote by $\operatorname{int} C$ the interior of a set $C \subset \mathbb{R}^n$.

2.3 Example (approximation of min-functions). *For* $f_i, f_i^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}$, $i = 1, \ldots, m$, *let*

$$f(x) = \min_{i=1,\ldots,m} f_i(x) \quad \textit{and} \quad f^\nu(x) = \min_{i=1,\ldots,m} f_i^\nu(x), \quad \forall x \in \mathbb{R}^n.$$

Then, $\operatorname{epi} f^\nu \xrightarrow{s} \operatorname{epi} f$ *whenever* $\operatorname{epi} f_i^\nu \xrightarrow{s} \operatorname{epi} f_i$ *for all* i.

Detail. Since $\operatorname{epi} f = \cup_{i=1}^m \operatorname{epi} f_i$ and likewise for $\operatorname{epi} f^\nu$, Proposition 2.2 establishes the conclusion. □

3 Minimization and Epi-Convergence

Set-convergence of epigraphs goes back to the pioneering work by Wijsman [30, 31] and Mosco [13] on convex functions. With an increasing focus on the nonconvex case in the late 1970s, Wets coined the term epi-convergence[6] [29]: For functions $f, f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}$, f^ν *epi-converges* to f, written $f^\nu \xrightarrow{e} f$, when the corresponding epigraphs set-converge, i.e.,

$$f^\nu \xrightarrow{e} f \quad \Longleftrightarrow \quad \operatorname{epi} f^\nu \xrightarrow{s} \operatorname{epi} f.$$

Since inner limits are contained in the corresponding outer limits, $\operatorname{epi} f^\nu \xrightarrow{s} \operatorname{epi} f$ takes place if and only if $\operatorname{LimOut}(\operatorname{epi} f^\nu) \subset \operatorname{epi} f \subset \operatorname{LimInn}(\operatorname{epi} f^\nu)$. The two inclusions translate into conditions (a) and (b) below; see [18, Proposition 7.2].

3.1 Theorem (characterization of epi-convergence). *For* $f, f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}$, $f^\nu \xrightarrow{e} f$ *if and only if*

(a) $\forall x^\nu \to x$, $\liminf f^\nu(x^\nu) \geq f(x)$,

(b) $\forall x$, $\exists\, x^\nu \to x$ *such that* $\limsup f^\nu(x^\nu) \leq f(x)$.

For (a), f^ν must be "high enough" regardless of how we approach x. For (b), there must be a "path" to x along which f^ν is "low enough". Figure 4 illustrates this with $f(x) = -1$ for $x \leq 0$ and $f(x) = 1$ otherwise, and $f^\nu(x) = -1$ for $x \leq -1/\nu$, $f^\nu(x) = -\sqrt{-\nu x}$ for $x \in (-1/\nu, 0]$, $f^\nu(x) = \sqrt{\nu x}$ for $x \in (0, 1/\nu]$, and $f^\nu(x) = 1$ otherwise. To verify epi-convergence, let's first consider (a). For any $x^\nu \to x \leq 0$, $f^\nu(x^\nu) \geq -1 = f(x)$ for all ν. For any $x^\nu \to x > 0$, $f^\nu(x^\nu) = 1 = f(x)$ for sufficiently large ν. Thus, (a) holds. Second, let's consider (b). For $x \neq 0$, we can take $x^\nu = x$ because $f^\nu(x^\nu) = f^\nu(x) \to f(x)$. It only remains to show that for some $x^\nu \to 0$, $\limsup f^\nu(x^\nu) \leq f(0) = -1$. Here, we need to be careful: x^ν above 0 or too close to 0 has $f^\nu(x^\nu)$ much above -1. However, $x^\nu = -1/\nu$ results in $f^\nu(x^\nu) = -1$ for all ν and we still have $x^\nu \to 0$ so (b) holds.

3.2 Example (weak convergence of distribution functions). *In probability theory, a sequence of distribution functions* $F^\nu : \mathbb{R}^n \to [0,1]$ *converge weakly to a distribution function* $F : \mathbb{R}^n \to [0,1]$ *when*

$$F^\nu(x) \to F(x) \textit{ for all } x \in \mathbb{R}^n \textit{ at which } F \textit{ is continuous.}$$

[6] Independently, in their work on calculus of variation, Di Giorgi and co-authors came to the same notion of set-convergence of epigraphs and called it Γ-convergence; see [9].

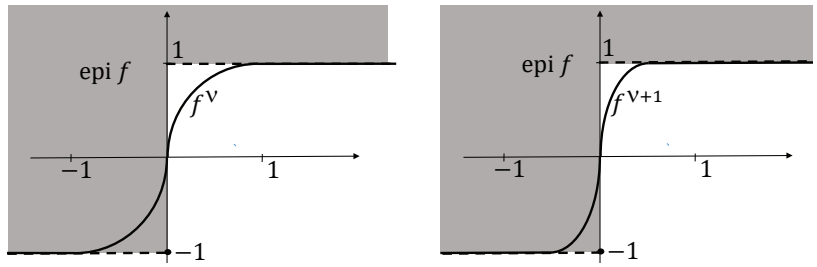

Figure 4: Functions f^ν (solid line) epi-converge to f (dashed line).

This is an example of set-convergence. Specifically, by [23, Theorem 3.2],

$$F^\nu(x) \text{ converges weakly to } F(x) \iff -F^\nu \xrightarrow{e} -F \iff \operatorname{hypo} F^\nu \xrightarrow{s} \operatorname{hypo} F,$$

where for any $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $\operatorname{hypo} f = \big\{(x,\alpha) \in \mathbb{R}^n \times \mathbb{R} \;\big|\; f(x) \geq \alpha\big\}$ *is the hypograph*[7] *of* f.

Detail. It's the monotonicity and upper-semicontinuity of distribution functions that allow us to make the equivalence between pointwise convergence and epi-convergence, which doesn't hold in general. If Figure 4 is modified by setting $f(0) = 1$ and f is replaced by $\frac{1}{2}f + \frac{1}{2}$ then the resulting function, say F, is a distribution function with a discontinuity at $x = 0$. Likewise, if we replace f^ν by $\frac{1}{2}f^\nu + \frac{1}{2}$, then the resulting function F^ν is a distribution function. For $x \neq 0$, $F^\nu(x) \to F(x)$ so F^ν converges weakly to F. Meanwhile, using a similar argument as in the discussion above, we obtain that $-F^\nu \xrightarrow{e} -F$. ◻

3.3 Example (density estimation using location mixtures). *Given a sample* $\xi^1, \ldots, \xi^n \in \mathbb{R}^m$, *the Kiefer-Wolfowitz maximum likelihood estimate of a probability density function on* $\mathbb{R}^m$, *used in clustering and denoising, is a minimizer of*

$$\underset{p \in \mathcal{P}}{\text{minimize}} -\frac{1}{n}\sum\nolimits_{j=1}^{n} \ln p(\xi^j),$$

where $\mathcal{P}$ *is the family of location mixtures of the* m*-variate standard normal density* $\varphi = (2\pi)^{-m/2}\exp(-\frac{1}{2}\|\cdot\|_2^2)$*; see [26]. The problem is equivalent to minimizing* $-\frac{1}{n}\sum_{j=1}^{n} \ln x_j$ *subject to*[8]

$$x \in C = \operatorname{con}\Big\{\big(\varphi(\xi^1 - z), \ldots, \varphi(\xi^n - z)\big) \;\Big|\; z \in \mathbb{R}^m\Big\}.$$

Since C *is the convex hull of an infinite number points, it needs to be approximated.*

Detail. Given $\bar{z}^1, \ldots, \bar{z}^\nu \in \mathbb{R}^m$, we can approximate C by

$$C^\nu = \operatorname{con}\Big\{\big(\varphi(\xi^1 - \bar{z}^k), \ldots, \varphi(\xi^n - \bar{z}^k)\big) \;\Big|\; k = 1, \ldots, \nu\Big\}.$$

[7] In general, set-convergence of hypographs is referred to as *hypo-convergence* with analogous properties to those of epi-convergence but with its primary application area being maximization instead of minimization.

[8] For $C \subset \mathbb{R}^n$, $\operatorname{con} C$ is its convex hull.

The approximating problem with C replaced by C^ν is tractable because of the equivalent convex problem

$$\underset{\mu\in\mathbb{R}^\nu}{\text{minimize}} -\frac{1}{n}\sum\nolimits_{j=1}^{n} \ln\Big(\sum\nolimits_{k=1}^{\nu} \mu_k\varphi(\xi^j-\bar{z}^k)\Big) \text{ s.t. } \sum\nolimits_{k=1}^{\nu}\mu_k=1,\ \mu_k\geq 0\ \forall k=1,\ldots,\nu.$$

To justify the approximation, let $f(x)=-\frac{1}{n}\sum_{j=1}^n \ln x_j+\iota_C(x)$ and $f^\nu(x)=-\frac{1}{n}\sum_{j=1}^n \ln x_j+\iota_{C^\nu}(x)$, which represent the actual and approximating problems, respectively. (Here, $-\ln\alpha=\infty$ for $\alpha\leq 0$.) We show that $f^\nu \xrightarrow{e} f$ when $\{\bar{z}^k, k\in\mathbb{N}\}$ has the property that for any $\bar{z}\in\mathbb{R}^m$, there exists $\bar{z}^k\to\bar{z}$.

First, consider Theorem 3.1(a) and let $x^\nu\to x$. Since $C^\nu\subset C$ and f is lsc, $\operatorname{liminf} f^\nu(x^\nu)\geq \operatorname{liminf} f(x^\nu)\geq f(x)$. Second, consider Theorem 3.1(b) and let $x\in C$. We need to construct $x^\nu\in C^\nu\to x$. By Caratheodory's Theorem, there exists $\alpha_0,\ldots,\alpha_n\geq 0$ and $z^0,\ldots,z^n\in\mathbb{R}^m$ such that $\sum_{j=0}^n\alpha_j=1$ and $x=\sum_{j=0}^n\alpha_j(\varphi(\xi^1-z^j),\ldots,\varphi(\xi^n-z^j))$. Set $x^\nu=\sum_{j=0}^n\alpha_j(\varphi(\xi^1-u_j^\nu),\ldots,\varphi(\xi^n-u_j^\nu))$, with $u_j^\nu\in\operatorname{argmin}_{k=1,\ldots,\nu}\|z^j-\bar{z}^k\|_2$. Certainly, $x^\nu\in C^\nu$. Since $\bar{z}^1,\bar{z}^2,\ldots$ eventually "fill" $\mathbb{R}^m$ by assumption, $u_j^\nu\to z^j$ as $\nu\to\infty$ for all j. By the continuity of φ, we have that $x^\nu\to x$ and also $f^\nu(x^\nu)\to f(x)$ by the continuity of the log-function. We have established the conditions of Theorem 3.1 and $f^\nu\xrightarrow{e} f$. ◻

For other applications of set-convergence in probability and statistics; see [17, 1] and §7.

Epi-convergence of composite function can often be established by checking properties of its various components. This might offer an easier path than working directly with Theorem 3.1.

3.4 Proposition (epi-convergence of sums). *For functions $f,f^\nu,g,g^\nu:\mathbb{R}^n\to(-\infty,\infty]$, suppose that $f^\nu\xrightarrow{e} f$ and $g^\nu\xrightarrow{e} g$. Then, $f^\nu+g^\nu\xrightarrow{e} f+g$ under either one of the following two conditions:*

(a) $-g^\nu\xrightarrow{e}-g$.

(b) $g^\nu(x)\to g(x)$ *and* $f^\nu(x)\to f(x)$ *for all* $x\in\mathbb{R}^n$.

The pointwise convergence of g^ν (f^ν) in (b) holds automatically if all g^ν (f^ν) are convex and real-valued.

Proof. For sequences $\{\alpha^\nu,\beta^\nu\in\overline{\mathbb{R}},\nu\in\mathbb{N}\}$, $\operatorname{liminf}(\alpha^\nu+\beta^\nu)\geq\operatorname{liminf}\alpha^\nu+\operatorname{liminf}\beta^\nu$ provided that the right-hand side is not $\infty-\infty$ or $-\infty+\infty$, and likewise for limsup but then with $\leq$. If $x^\nu\to x$, then

$$\operatorname{liminf}\big(f^\nu(x^\nu)+g^\nu(x^\nu)\big)\geq\operatorname{liminf}f^\nu(x^\nu)+\operatorname{liminf}g^\nu(x^\nu)\geq f(x)+g(x)$$

and Theorem 3.1(a) holds without either condition (a) or (b). Given $x\in\mathbb{R}^n$, there exists $x^\nu\to x$ such that $\operatorname{limsup}f^\nu(x^\nu)\leq f(x)$. We note that $\operatorname{limsup}f^\nu(x^\nu)>-\infty$ because otherwise $\operatorname{liminf}f^\nu(x^\nu)=-\infty$ and $f(x)=-\infty$, but this violates the fact that $f>-\infty$. Under (a), $g^\nu(x^\nu)\to g(x)$ and

$$\operatorname{limsup}\big(f^\nu(x^\nu)+g^\nu(x^\nu)\big)\leq\operatorname{limsup}f^\nu(x^\nu)+\operatorname{limsup}g^\nu(x^\nu)\leq f(x)+g(x)$$

and Theorem 3.1(b) holds. Under condition (b), set $x^\nu=x$ so that

$$\operatorname{limsup}\big(f^\nu(x^\nu)+g^\nu(x^\nu)\big)\leq\operatorname{limsup}f^\nu(x)+\operatorname{limsup}g^\nu(x)=f(x)+g(x)$$

and Theorem 3.1(b) holds again.

When all the functions g^ν are convex and real-valued, then $g^\nu\xrightarrow{e} g$ if and only if $g^\nu(x)\to g(x)$ for all $x\in\mathbb{R}^n$ by [18, Theorem 7.17]. Thus, (b) holds automatically under the additional assumption. ◻

We note that having $g^\nu \xrightarrow{e} g$ and $-g^\nu \xrightarrow{e} -g$ is equivalent to having $g^\nu(x^\nu) \to g(x)$ for all $x^\nu \to x$. Thus, condition (a) of the proposition holds, for example, when $g^\nu = g$ and g is continuous. It is now possible to verify the claim in the Introduction regarding epi-convergence for approximations obtained by constraint softening using Theorem 3.1 and Proposition 3.4.

3.5 Proposition (epi-convergence of composite functions). *For $h, h^\nu : \mathbb{R} \to \overline{\mathbb{R}}$ and $f, f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}$, the following hold:*

(a) *If $h^\nu \xrightarrow{e} h$, $h^\nu(\alpha) \to h(\alpha)$ for all $\alpha \in \mathbb{R}$, and f is real-valued and continuous, then $h^\nu \circ f \xrightarrow{e} h \circ f$.*

(b) *If $f^\nu \xrightarrow{e} f$ and h is continuous, nondecreasing and extended to $\overline{\mathbb{R}}$ with the conventions $h(-\infty) = \inf_{\alpha\in\mathbb{R}} h(\alpha)$ and $h(\infty) = \sup_{\alpha\in\mathbb{R}} h(\alpha)$, then $h \circ f^\nu \xrightarrow{e} h \circ f$.*

Proof. For (a), let $x^\nu \to x$, which ensures that $f(x^\nu) \to f(x)$. By Theorem 3.1(a), $\liminf h^\nu(f(x^\nu)) \geq h(f(x))$. Given $x \in \mathbb{R}^n$, set $x^\nu = x$. Then, $\limsup h^\nu(f(x^\nu)) = \limsup h^\nu(f(x)) = h(f(x))$. We have confirmed both conditions in Theorem 3.1.

For (b), let $x^\nu \to x$, which implies that $\liminf f^\nu(x^\nu) \geq f(x)$ by Theorem 3.1(a). Fix ν and let $\varepsilon > 0$. Suppose that $\alpha^\nu = \inf_{\mu\geq\nu} h(f^\mu(x^\mu)) \in \mathbb{R}$. Then, there exists $\bar{\mu} \geq \nu$ such that $\alpha^\nu \geq h(f^{\bar{\mu}}(x^{\bar{\mu}})) - \varepsilon \geq h(\inf_{\mu\geq\nu} f^\mu(x^\mu)) - \varepsilon$, the last inequality holds because h is nondecreasing. Since ε is arbitrary, $\alpha^\nu \geq h(\inf_{\mu\geq\nu} f^\mu(x^\mu))$. A similar argument leads to the same inequality if $\alpha^\nu = -\infty$ and, trivially, also when $\alpha^\nu = \infty$. Since the inequality holds for all ν, it follows by the continuity of h that

$$\begin{aligned}\liminf h(f^\nu(x^\nu)) &= \lim_{\nu\to\infty} \big(\inf_{\mu\geq\nu} h(f^\mu(x^\mu))\big) \geq \lim_{\nu\to\infty} h\big(\inf_{\mu\geq\nu} f^\mu(x^\mu)\big)\\ &= h\big(\lim_{\nu\to\infty}(\inf_{\mu\geq\nu} f^\mu(x^\mu))\big) = h\big(\liminf f^\nu(x^\nu)\big) \geq h(f(x)).\end{aligned}$$

Given $x \in \mathbb{R}^n$, there exists $x^\nu \to x$ such that $f^\nu(x^\nu) \to f(x)$ by Theorem 3.1(b). Since h is continuous, this implies $h(f^\nu(x^\nu)) \to h(f(x))$ and in view Theorem 3.1 we have established (b). □

As indicated in the Introduction, $f^\nu \xrightarrow{e} f$ ensures that cluster points of minimizers of f^ν are minimizers of f but the situation for minimum values is more delicate. If $f^\nu(x) = \max\{-1, x/\nu\}$ and $f(x) = 0$ for all x, then $f^\nu \xrightarrow{e} f$ but $\inf f^\nu = -1 < \inf f = 0$. The trouble is that $\mathrm{Lim}(\operatorname{argmin} f^\nu) = \emptyset$. This pathological situation is eliminated under *tightness.*

3.6 Definition (tightness). *The functions $\{f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}, \nu \in \mathbb{N}\}$ are tight if for all $\varepsilon > 0$, there is a compact $B_\varepsilon \subset \mathbb{R}^n$ and an index ν_ε such that*

$$\inf_{x\in B_\varepsilon} f^\nu(x) \leq \inf f^\nu + \varepsilon \text{ for all } \nu \geq \nu_\varepsilon.$$

The functions epi-converge tightly *if in addition to being tight they also epi-converge to some function.*

Tightness holds, for example, if $\{\operatorname{dom} f^\nu, \nu \in \mathbb{N}\}$ are contained in a bounded set or if there is a bounded set $B \subset \mathbb{R}^n$ such that $B \cap \operatorname{argmin} f^\nu$ is nonempty for all ν. With the refinement of tightness, we obtain the following consequences of epi-convergence (see [24, Theorem 3.8]), where

$$\varepsilon\text{-}\operatorname{argmin} f = \{x \in \operatorname{dom} f \mid f(x) \leq \inf f + \varepsilon\} \text{ for } f : \mathbb{R}^n \to \overline{\mathbb{R}} \text{ and } \varepsilon \in \mathbb{R}_+ = [0, \infty).$$

3.7 Theorem (consequences of epi-convergence). *For proper $f, f^\nu : \mathbb{R}^n \to \overline{\mathbb{R}}$, suppose that $f^\nu \xrightarrow{e} f$. Then, the following hold:*

(a) $\forall\{\varepsilon^\nu \in \mathbb{R}_+ \to 0, \nu \in \mathbb{N}\}$, $\operatorname{LimOut}\left(\varepsilon^\nu\text{-}\operatorname{argmin} f^\nu\right) \subset \operatorname{argmin} f$.

(b) $\operatorname{limsup}(\inf f^\nu) \leq \inf f$.

(c) *If* $\{x^\nu \in \operatorname{argmin} f^\nu, \nu \in N\}$ *converges for some* $N \in \mathcal{N}_\infty^\#$, *then* $\lim_{\nu \in N}(\inf f^\nu) = \inf f$.

(d) $\inf f^\nu \to \inf f > -\infty \iff \{f^\nu, \nu \in \mathbb{N}\}$ *is tight.*

(e) $\inf f^\nu \to \inf f$ *and* $\varepsilon > 0 \implies \operatorname{LimInn}\left(\varepsilon\text{-}\operatorname{argmin} f^\nu\right) \supset \operatorname{argmin} f$.

(f) $\inf f^\nu \to \inf f \implies \exists\{\varepsilon^\nu \in \mathbb{R}_+ \to 0, \nu \in \mathbb{N}\}$ *such that* $\operatorname{Lim}\left(\varepsilon^\nu\text{-}\operatorname{argmin} f^\nu\right) = \operatorname{argmin} f$.

(g) $\exists\{\varepsilon^\nu \in \mathbb{R}_+ \to 0, \nu \in \mathbb{N}\}$ *such that* $\operatorname{Lim}\left(\varepsilon^\nu\text{-}\operatorname{argmin} f^\nu\right) = \operatorname{argmin} f \neq \emptyset \implies \inf f^\nu \to \inf f > -\infty$.

Part (a) formalizes the earlier assertion about minimizers. In light of the discussion prior to the theorem, (b,c) is as much as one can hope to say about minimum values without tightness. Part (f) supplements (a) by establishing set-convergence to $\operatorname{argmin} f$ when the tolerance ε^ν vanishes slowly enough.

4 Variational Geometry and Subdifferentiability

Normal and tangent spaces are well-known from Differential Geometry as means to locally characterize sets defined by smooth manifolds. However, they fall short of the demand of variational problems where nonsmoothness is prevalent. Set-convergence supports the development of extensions of these concepts. This brings us to normal cones and tangent cones.

The *tangent cone*[9] to $C \subset \mathbb{R}^n$ at one of its points $\bar{x}$ is defined as[10]

$$T_C(\bar{x}) = \operatorname{LimOut} \nu(C - \bar{x}), \text{ with } w \in T_C(\bar{x}) \text{ being a } \textit{tangent vector} \text{ to } C \text{ at } \bar{x}.$$

Geometrically, $C - \bar{x}$ is a translation of C to the origin such that 0 occupies the same relative position in $C - \bar{x}$ as $\bar{x}$ in C and $T_C(\bar{x})$ is the outer limit of this translation, "magnified" by ν; see Figure 5.

Figure 6 illustrates typical situations. In the left portion, C has a "smooth" boundary at $\bar{x}$ and $T_C(\bar{x})$ is simply the halfspace formed by the usual tangent space of the boundary of C at $\bar{x}$. When $\bar{x}$ is at a "corner," as seen in the middle portion of the figure, the tangent cone narrows. At the "inward kink" of the heart-shaped set to the right, the tangent cone fans out in all directions except some pointing upward. In this figure and elsewhere, tangent vectors are often visualized after being translated by $\bar{x}$, i.e., the beginning of an arrow visualizing a tangent vector w is moved from the origin, where it always will be, to $\bar{x}$ so the illustrations are actually of $\bar{x} + w$ and not w.

For $\bar{x} \in C \subset \mathbb{R}^n$, $v \in \mathbb{R}^n$ is a *regular normal vector* to C at $\bar{x}$ if

$$\langle v, w \rangle \leq 0 \text{ for all } w \in T_C(\bar{x}).$$

[9] A tangent cone is sometimes called a contingent cone.

[10] For $C \subset \mathbb{R}^n$, $\lambda \in \mathbb{R}$, and $a \in \mathbb{R}^n$: $\lambda C = \{\lambda x \mid x \in C\}$ (scalar multiplication) and $C + a = \{x + a \mid x \in C\}$ (addition).

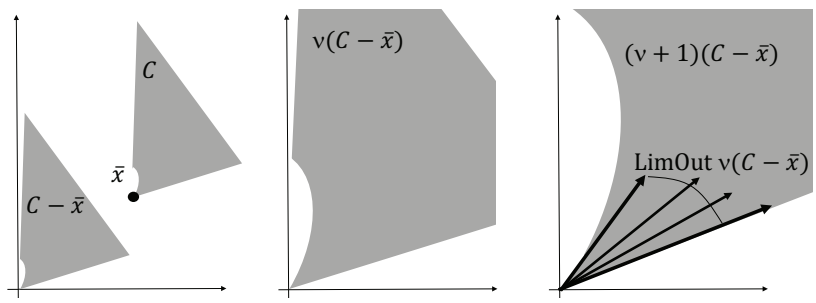

Figure 5: Translation (left), magnification (middle), further magnification and tangent cone (right).

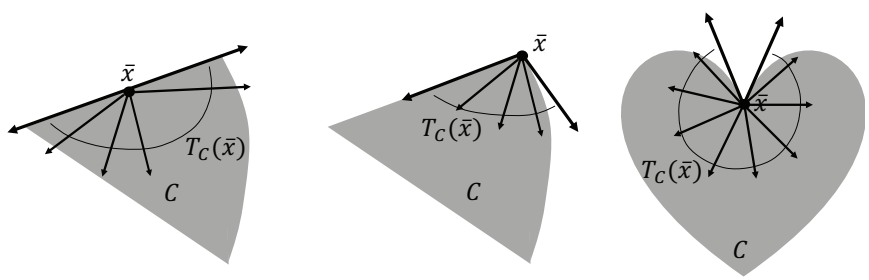

Figure 6: Tangent vectors and tangent cones translated to $\bar{x}$.

The set of all such regular normal vectors is $\widehat{N}_C(\bar{x})$. The *normal cone* to C at $\bar{x}$ is defined as

$$N_C(\bar{x}) = \bigcup_{x^\nu \in C \to \bar{x}} \text{LimOut}\, \widehat{N}_C(x^\nu), \quad \text{with } v \in N_C(\bar{x}) \text{ being a } \textit{normal vector} \text{ to } C \text{ at } \bar{x}.$$

We recall that $\langle v, w \rangle = \|v\|_2 \|w\|_2 \cos\alpha$, where α is the angle between the vectors v and w. Thus, a regular normal vector forms an angle of between 90 and 270 degrees with every tangent vector and thereby extends the "perpendicular relationship" of normal and tangent spaces from Differential Geometry. Figure 7 illustrates the normal cones corresponding to the tangent cones of Figure 6.

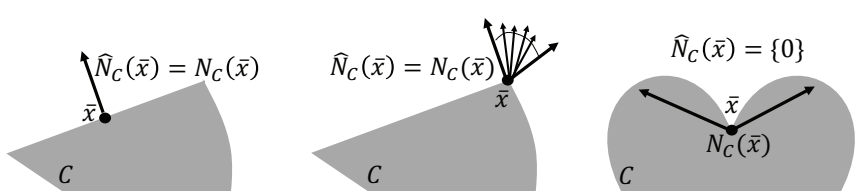

Figure 7: Normal vectors and normal cones translated to $\bar{x}$.

On their own, regular normal vectors fail to provide a solid basis for the treatment of "complicated" sets as seen in the right portion of Figure 7. At the inward kink in this heart-shaped set, $\widehat{N}_C(\bar{x}) = \{0\}$ since $T_C(\bar{x})$ fans out in nearly all directions and it becomes impossible to find a nonzero vector that

forms an angle of at least 90 degrees with all tangent vectors. We obtain a robust notion of "normality" by considering the union of the outer limits of $\widehat{N}_C(x^\nu)$ for all $x^\nu \in C \to \bar{x}$. This leads to vectors in $N_C(\bar{x})$ that aren't regular normals. Figure 8 shows $v^\nu \in \widehat{N}_C(x^\nu)$ at points $x^\nu \in C$ approaching $\bar{x}$ from the left. In the limit, these regular normal vectors give rise to the normal vectors pointing northeast in the right-most portion of the figure. Likewise, $x^\nu \in C$ approaching $\bar{x}$ from the right result in the normal vectors at $\bar{x}$ pointing northwest. It might be counterintuitive that the normal vectors at $\bar{x}$ points *into* the set, but this is indeed representative of the local geometry at this point and allows us to treat such points in a meaningful way. In general, a normal cone is a closed set.

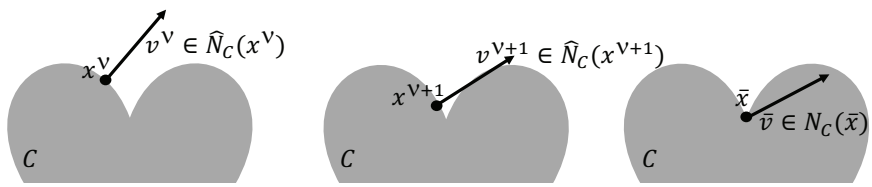

Figure 8: Normal vectors as limits of regular normal vectors.

4.1 Example (optimality conditions). *For a smooth*[11] *function $f : \mathbb{R}^n \to \mathbb{R}$ and a closed set $C \subset \mathbb{R}^n$, consider the problem of minimizing f over C. Then,*

$$x^\star \text{ is a local minimizer} \quad \Longrightarrow \quad -\nabla f(x^\star) \in N_C(x^\star),$$

which can be seen, for example, by invoking [18, Exercise 10.10]. Suppose that a numerical procedure generates the sequence $\{x^\nu \in C, \nu \in \mathbb{N}\}$ with $\operatorname{dist}_2(-\nabla f(x^\nu), N_C(x^\nu)) \le \varepsilon^\nu$ for some tolerance $\varepsilon^\nu \searrow 0$. Then, the continuity property built into normal cones ensures that any cluster point $\bar{x}$ of the sequence satisfies $-\nabla f(\bar{x}) \in N_C(\bar{x})$.

Detail. Suppose that $x^\nu \xrightarrow[N]{} \bar{x}$ for some $N \in \mathcal{N}_\infty^\#$. By continuity, $\nabla f(x^\nu) \xrightarrow[N]{} \nabla f(\bar{x})$. Let $v^\nu \in \operatorname{argmin}\{\|v - \nabla f(x^\nu)\|_2 \mid -v \in N_C(x^\nu)\}$. Then, $\|v^\nu - \nabla f(x^\nu)\|_2 \le \varepsilon^\nu$, which implies that $v^\nu \xrightarrow[N]{} \nabla f(\bar{x})$. Since $-v^\nu \in N_C(x^\nu) \xrightarrow[N]{} -\nabla f(\bar{x})$, we must have $-\nabla f(\bar{x}) \in \operatorname{LimOut}_{\nu \in N} N_C(x^\nu)$. From the definition of normal cones, $\operatorname{LimOut}_{\nu \in N} N_C(x^\nu) \subset N_C(\bar{x})$ because $x^\nu \xrightarrow[N]{} \bar{x}$ and the assertion follows. ◻

We next turn the attention to subgradients, which are defined in terms of normal cones of epigraphs. For $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\bar{x}$ with $f(\bar{x})$ finite, v is a *subgradient*[12] of f at $\bar{x}$ if

$$(v, -1) \in N_{\operatorname{epi} f}\big(\bar{x}, f(\bar{x})\big).$$

The set of all subgradients of f at $\bar{x}$ is denoted by $\partial f(\bar{x})$.

[11] A function is smooth if its gradient is continuous.

[12] These subgradients are also called Mordukhovich or limiting subgradients, to distinguish them from other kinds, but we omit such qualifications because, at least in finite dimensions, they are the central ones; see [18, Chapter 8]. For convex functions, these subgradients coincide with those from convex analysis.

As an introductory example, let $f(x) = x^2$ for $x \in \mathbb{R}$. Then, $N_{\operatorname{epi} f}(1, f(1)) = \{(2, -1)y \mid y \geq 0\}$. Thus, $(v, -1) \in N_{\operatorname{epi} f}(1, f(1))$ if and only if $v = 2$ and $\partial f(1) = \{2\}$ as expected. Figure 9 illustrates the normal cones of an epigraph at two points. Roughly, it seems like $N_{\operatorname{epi} f}(\bar{x}, f(\bar{x}))$ contains vectors of the form $\lambda(v, -1)$ for $\lambda \geq 0$ and $v \in [1/2, 2]$; these are the subgradients of f at $\bar{x}$. Note that 2 is the slope of f to the right of $\bar{x}$ and $1/2$ is the slope to the left. At $(x', f(x'))$, normal vectors appear to be of the form $\lambda(-4, -1)$ and $\lambda(-1/2, -1)$ for $\lambda \geq 0$. Then, $v = -4$ and $v = -1/2$ are subgradients and in fact the only ones because the normal cone at $(x', f(x'))$ consists of only two rays.

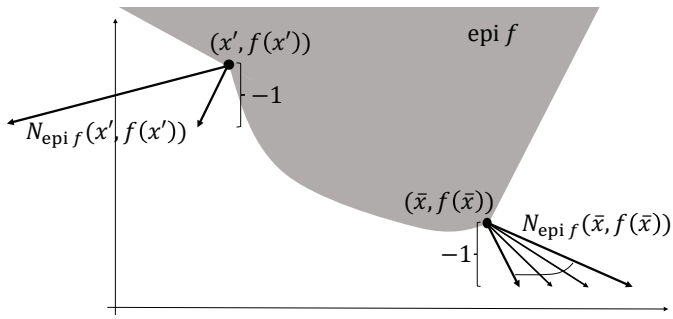

Figure 9: Normal cones of epi f.

4.2 Example (the Fermat rule). *For $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $x^\star$ with $f(x^\star) > -\infty$,*

$$x^\star \text{ is a local minimizer of } f \implies 0 \in \partial f(x^\star),$$

which is the celebrated Fermat rule; see [18, Theorem 8.15]. If an algorithm generates a sequence $\{x^\nu, \nu \in \mathbb{N}\}$ with $\operatorname{dist}_2(0, \partial f(x^\nu)) \leq \varepsilon^\nu$ for some tolerance $\varepsilon^\nu \searrow 0$, $\bar{x}$ is a cluster point, i.e., $x^\nu \xrightarrow[N]{} \bar{x}$ for some $N \in \mathcal{N}_\infty^\#$, and $f(x^\nu) \xrightarrow[N]{} f(\bar{x}) \in \mathbb{R}$, then $\bar{x}$ satisfies the Fermat Rule, i.e, $0 \in \partial f(\bar{x})$.

Detail. In light of Example 4.1 and the definition of subgradients, $\operatorname{LimOut}_{\nu \in N} \partial f(x^\nu) \subset \partial f(\bar{x})$; see also [18, Proposition 8.7]. Arguments similar to the ones in Example 4.1 establish the assertion. ◻

These concepts are supplemented by subderivatives of functions (akin to directional derivatives) which are defined in terms of outer limits of epigraphs of certain difference quotients, with further extensions leading to notions of differentiation of set-valued mappings again utilizing set-convergence as a primary tool; see [18, Chapter 8]. We refer to the commentary sections of [18, Chapters 6 and 8] for the historical development of these concepts.

5 Generalized Equations and Graphical Convergence

While we express a system of equations $F(x) = 0$ in terms of a mapping $F : \mathbb{R}^n \to \mathbb{R}^m$, a generalized equation is given by a *set-valued mapping* $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$. The distinguishing feature of S relative to F is

that we now allow for an input x to produce an output $S(x)$ that is a subset of $\mathbb{R}^m$ and not necessarily a single point. In fact, we even allow the subset to be empty so that the domain

$$\operatorname{dom} S = \big\{x \in \mathbb{R}^n \;\big|\; S(x) \neq \emptyset\big\}$$

could be a strict subset of $\mathbb{R}^n$. Of course, F defines a particular set-valued mapping with $S(x) = \{F(x)\}$ for all $x \in \mathbb{R}^n$ with $\operatorname{dom} S = \mathbb{R}^n$. The *graph* of a set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is the set

$$\operatorname{gph} S = \big\{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \;\big|\; y \in S(x)\big\},$$

which describes S fully; see Figure 10. Given $\bar{y} \in \mathbb{R}^m$ and $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, a *generalized equation* is then

$$\bar{y} \in S(x), \quad \text{with set of solutions } S^{-1}(\bar{y}) = \big\{x \in \mathbb{R}^n \;\big|\; \bar{y} \in S(x)\big\}.$$

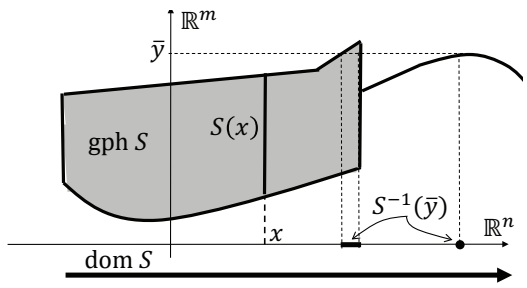

Figure 10: The solution set $S^{-1}(\bar{y})$ to the generalized equation $\bar{y} \in S(x)$.

The sets of subgradients of $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ can be viewed as a set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ with $S(x) = \partial f(x)$ if $f(x)$ is finite and $S(x) = \emptyset$ otherwise. Likewise, the normal cones of $C \subset \mathbb{R}^n$ define a set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ with $S(x) = N_C(x)$ if $x \in C$ and $S(x) = \emptyset$ otherwise. The facts about $\operatorname{LimOut} N_C(x^\nu) \subset N_C(\bar{x})$ and $\operatorname{LimOut} \partial f(x^\nu) \subset \partial f(\bar{x})$ that we utilize in Examples 4.1 and 4.2 can be viewed as continuity properties of the corresponding set-valued mappings.

Generally, a set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is *outer semicontinuous* (osc) at $\bar{x} \in \mathbb{R}^n$ if

$$\bigcup_{x^\nu \to \bar{x}} \operatorname{LimOut} S(x^\nu) \subset S(\bar{x}).$$

It's osc if this holds for all $\bar{x} \in \mathbb{R}^n$. Figure 10 furnishes an example.

5.1 Proposition (characterization of osc). *For $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$,*

$$\operatorname{gph} S \text{ is closed} \iff S \text{ is osc} \iff S^{-1} \text{ is osc.}$$

Proof. A closed $\operatorname{gph} S$ means that every sequence $(x^\nu, v^\nu) \in \operatorname{gph} S \to (\bar{x}, \bar{v})$ has $(\bar{x}, \bar{v}) \in \operatorname{gph} S$. But, this is equivalent to $x^\nu \to \bar{x}$, $v^\nu \to \bar{v}$, and $v^\nu \in S(x^\nu)$ implying $\bar{v} \in S(\bar{x})$, which is guaranteed by osc. The claim about the inverse mapping follows similarly because $v \in S(x)$ if and only if $x \in S^{-1}(v)$. $\square$

5.2 Example (osc of feasible-set mapping). *For a closed $X \subset \mathbb{R}^n$ and continuous $f_i, g_i : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$, consider the set-valued mapping $S : \mathbb{R}^d \rightrightarrows \mathbb{R}^n$ given by*

$$S(u) = \big\{x \in X \;\big|\; f_i(u,x) = 0, i = 1,\dots,m; \quad g_i(u,x) \le 0, i = 1,\dots,q\big\},$$

which arises in stability studies of feasible sets under perturbations. In fact, S is osc.

Detail. In view of Proposition 5.1, it suffices to establish the closedness of the graph

$$\operatorname{gph} S = \big\{(u,x) \in \mathbb{R}^d \times \mathbb{R}^n \;\big|\; x \in S(u)\big\} = \bigcap_{i=1}^{m} \big\{(u,x)\big|f_i(u,x) = 0\big\} \bigcap_{i=1}^{q} \big\{(u,x)\big|g_i(u,x) \le 0\big\} \bigcap \big(\mathbb{R}^d \times X\big).$$

Since X is closed and the functions involved are continuous, this is an intersection of closed sets and is therefore closed. ◻

A set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is *inner semicontinuous* (isc) at $\bar{x} \in \mathbb{R}^n$ if

$$\bigcap_{x^\nu \to \bar{x}} \operatorname{LimInn} S(x^\nu) \supset S(\bar{x}).$$

It's isc if this holds for all $\bar{x} \in \mathbb{R}^n$. Moreover, S is *continuous* (at $\bar{x}$) if it's both osc and isc (at $\bar{x}$), in which case $S(x^\nu) \xrightarrow{s} S(\bar{x})$ for all $x^\nu \to \bar{x}$.

While osc is common in applications, isc occurs less frequently. As an example, if $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ has a convex graph, then S is isc at any point $\bar{x} \in \operatorname{int}(\operatorname{dom} S)$; see [18, Theorem 5.9b].

Applications often bring us generalized equations that can't be solved by existing algorithms, at least not in reasonable time, and it becomes necessary to consider approximations. Approximations may also be introduced artificially in sensitivity analysis. Regardless of the circumstances, we would like to examine whether replacing $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ by an alternative $S^\nu : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ results in significantly different solutions of the corresponding generalized equations. It's immediately clear that if an algorithm computes $\{x^\nu, \nu \in \mathbb{N}\}$ with the property that $\operatorname{dist}_2(\bar{y}, S^\nu(x^\nu)) \le \varepsilon^\nu$ for some vanishing tolerances $\varepsilon^\nu \ge 0$, then every cluster point $\bar{x}$ of the sequence satisfies $\bar{y} \in S(\bar{x})$ provided that

$$\operatorname{LimOut}(\operatorname{gph} S^\nu) \subset \operatorname{gph} S.$$

To see this, note that the vanishing $\operatorname{dist}_2(\bar{y}, S^\nu(x^\nu))$ guarantees that there is $y^\nu \in S^\nu(x^\nu)$ such that $y^\nu \to \bar{y}$. Since $(x^\nu, y^\nu) \in \operatorname{gph} S^\nu$, it follows by the definition of outer limits that $(\bar{x}, \bar{y}) \in \operatorname{LimOut}(\operatorname{gph} S^\nu)$ and then also $(\bar{x}, \bar{y}) \in \operatorname{gph} S$ under the stated assumption. Consequently, the algorithm is justified.

5.3 Example (smoothing of complementarity problems). *For a smooth*[13] *mapping $F : \mathbb{R}^n \to \mathbb{R}^n$, the complementarity problem*

$$x \ge 0, \quad F(x) \ge 0, \quad \big\langle F(x), x\big\rangle = 0$$

corresponds to the generalized equation $0 \in F(x) + N_C(x)$, with $C = \mathbb{R}^n_+$. This, in turn, is equivalent to the normal map condition

$$0 = F\big(\operatorname{prj}_C(z)\big) + z - \operatorname{prj}_C(z),$$

[13] A mapping is smooth if its Jacobian is continuous.

where the projection $\operatorname{prj}_C(z) = (\max\{0, z_1\}, \ldots, \max\{0, z_n\})$*; see [16]. Although just "regular" equations, the normal map condition involves nonsmooth functions and the usual Newton's Method doesn't apply. However, a smooth approximation can be brought in.*

Detail. For $\theta^\nu > 0$, $\varphi^\nu(\alpha) = \frac{1}{\theta^\nu} \ln\big(1 + e^{\alpha\theta^\nu}\big)$ furnishes a smooth approximation of $\alpha \mapsto \max\{0, \alpha\}$ with error

$$0 \le \varphi^\nu(\alpha) - \max\{0, \alpha\} \le \frac{\ln 2}{\theta^\nu} \quad \text{for all } \alpha \in \mathbb{R}.$$

Consequently, the projection mapping can be approximated by the smooth mapping $\Phi^\nu : \mathbb{R}^n \to \mathbb{R}$ given by $\Phi^\nu(z) = (\varphi^\nu(z_1), \ldots, \varphi^\nu(z_n))$. This results in the approximating equations

$$0 = F\big(\Phi^\nu(z)\big) + z - \Phi^\nu(z),$$

which can be solved using Newton's Method. We justify this approach by viewing the equations in terms of the set-valued mappings $S, S^\nu : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ given by $S(z) = \{F(\operatorname{prj}_C(z)) + z - \operatorname{prj}_C(z)\}$ and $S^\nu(z) = \{F(\Phi^\nu(z)) + z - \Phi^\nu(z)\}$ for all $z \in \mathbb{R}^n$. Clearly, the normal map condition is then $0 \in S(z)$. In view of the discussion prior to the example, we only need to establish that $\operatorname{LimOut}(\operatorname{gph} S^\nu) \subset \operatorname{gph} S$.

Suppose that $(\bar{z}, \bar{v}) \in \operatorname{LimOut}(\operatorname{gph} S^\nu)$. Then, there are $N \in \mathcal{N}_\infty^\#$, $z^\nu \underset{N}{\to} \bar{z}$, and $v^\nu \underset{N}{\to} \bar{v}$ such that $v^\nu \in S^\nu(z^\nu)$ for $\nu \in N$. Moreover,

$$\begin{aligned}\big\|F\big(\operatorname{prj}_C(\bar{z})\big) + \bar{z} - \operatorname{prj}_C(\bar{z}) - \bar{v}\big\|_2 \le{}& \big\|F\big(\operatorname{prj}_C(\bar{z})\big) + \bar{z} - \operatorname{prj}_C(\bar{z}) - F\big(\Phi^\nu(z^\nu)\big) - z^\nu + \Phi^\nu(z^\nu)\big\|_2 \\ &+ \big\|F\big(\Phi^\nu(z^\nu)\big) + z^\nu - \Phi^\nu(z^\nu) - \bar{v}\big\|_2.\end{aligned}$$

Since $F(\Phi^\nu(z^\nu)) + z^\nu - \Phi^\nu(z^\nu) = v^\nu \underset{N}{\to} \bar{v}$, the second term on the right-hand side vanishes. Also,

$$\begin{aligned}&\big\|F\big(\operatorname{prj}_C(\bar{z})\big) + \bar{z} - \operatorname{prj}_C(\bar{z}) - F\big(\Phi^\nu(z^\nu)\big) - z^\nu + \Phi^\nu(z^\nu)\big\|_2 \\ \le{}&\big\|F\big(\operatorname{prj}_C(\bar{z})\big) - F\big(\Phi^\nu(z^\nu)\big)\big\|_2 + \|\bar{z} - z^\nu\|_2 + \big\|\operatorname{prj}_C(\bar{z}) - \Phi^\nu(z^\nu)\big\|_2,\end{aligned}$$

which vanishes as long as $\theta^\nu \to \infty$ because then $\varphi^\nu(z_j^\nu) \underset{N}{\to} \max\{0, \bar{z}_j\}$ for all $j = 1, \ldots, n$. Consequently, $F(\operatorname{prj}_C(\bar{z})) + \bar{z} - \operatorname{prj}_C(\bar{z}) = \bar{v}$ and $(\bar{z}, \bar{v}) \in \operatorname{gph} S$. ◻

With the pioneering work of Spagnolo [27], Attouch [2], De Giorgi [8], and Moreau [12], set-convergence of graphs of set-valued mappings, which became known as graphical convergence, and its importance for generalized equations came to the foreground. For $S^\nu, S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, we say that S^ν *converges graphically* to S, written $S^\nu \xrightarrow{g} S$, when

$$\operatorname{gph} S^\nu \xrightarrow{s} \operatorname{gph} S.$$

In light of the definitions, this takes place when the condition $\operatorname{LimOut}(\operatorname{gph} S^\nu) \subset \operatorname{gph} S$ is supplemented by $\operatorname{LimInn}(\operatorname{gph} S^\nu) \supset \operatorname{gph} S$; see Figure 11.

Since $v \in S(x)$ if and only if $x \in S^{-1}(v)$, graphical convergence of set-valued mappings corresponds to graphical convergence of the corresponding inverse mappings:

$$S^\nu \xrightarrow{g} S \iff (S^\nu)^{-1} \xrightarrow{g} S^{-1}.$$

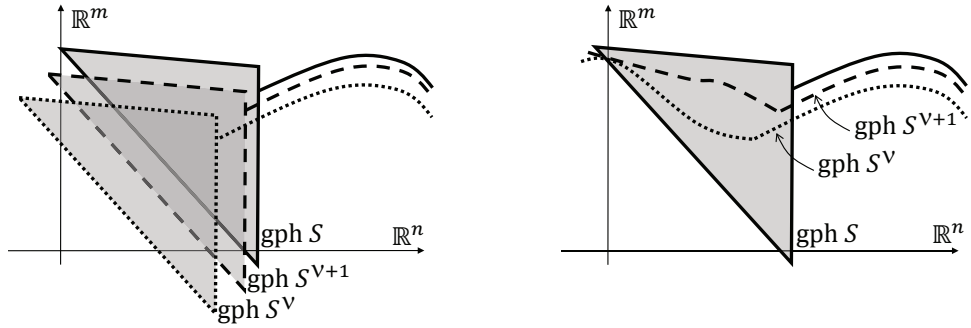

Figure 11: On the left, $\operatorname{gph} S^\nu \xrightarrow{s} \operatorname{gph} S$, but on the right only $\operatorname{LimOut}(\operatorname{gph} S^\nu) \subset \operatorname{gph} S$ holds.

Under graphical convergence, conclusions about solutions of approximating generalized equations can be strengthen beyond the discussion prior to Example 5.3. For $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, $\bar{y} \in \mathbb{R}$, and $\varepsilon \geq 0$, the *set of ε-solutions* to the generalized equation $\bar{y} \in S(x)$ is defined as

$$S^{-1}\big(\mathbb{B}(\bar{y},\varepsilon)\big) = \bigcup_{y \in \mathbb{B}(\bar{y},\varepsilon)} S^{-1}(y),$$

where the ball $\mathbb{B}(\bar{y},\varepsilon) = \{y \in \mathbb{R}^m \mid \|y - \bar{y}\| \leq \varepsilon\}$ is given in terms of any norm $\|\cdot\|$ on $\mathbb{R}^m$.

5.4 Theorem (consequences of graphical convergence). *For set-valued mappings $S, S^\nu : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, suppose that $S^\nu \xrightarrow{g} S$ and $\bar{y} \in \mathbb{R}^m$. Then, the following hold:*

(a) $\forall\{\varepsilon^\nu \in \mathbb{R}_+ \to 0, \nu \in \mathbb{N}\}$, $\operatorname{LimOut}(S^\nu)^{-1}\big(\mathbb{B}(\bar{y},\varepsilon^\nu)\big) \subset S^{-1}(\bar{y})$.

(b) $\exists\{\varepsilon^\nu \in \mathbb{R}_+ \to 0, \nu \in \mathbb{N}\}$, *with* $\operatorname{LimInn}(S^\nu)^{-1}\big(\mathbb{B}(\bar{y},\varepsilon^\nu)\big) \supset S^{-1}(\bar{y})$.

Proof. For (a), we only utilize $\operatorname{LimOut}(\operatorname{gph}(S^\nu)^{-1}) \subset \operatorname{gph} S^{-1}$: Let $\bar{x} \in \operatorname{LimOut}(S^\nu)^{-1}(\mathbb{B}(\bar{y},\varepsilon^\nu))$. Then, there exist $N \in \mathcal{N}_\infty^\#$ and $x^\nu \xrightarrow[N]{} \bar{x}$ such that $x^\nu \in (S^\nu)^{-1}(\mathbb{B}(\bar{y},\varepsilon^\nu))$ and, for some $y^\nu \in \mathbb{B}(\bar{y},\varepsilon^\nu)$, $(y^\nu, x^\nu) \in \operatorname{gph}(S^\nu)^{-1}$. Since cluster points of such sequences are contained in $\operatorname{gph} S^{-1}$, $\bar{x} \in S^{-1}(\bar{y})$.

For (b), we use $\operatorname{LimInn}(\operatorname{gph}(S^\nu)^{-1}) \supset \operatorname{gph} S^{-1}$: Let $\bar{x} \in S^{-1}(\bar{y})$. Then, $(\bar{y},\bar{x}) \in \operatorname{LimInn}(\operatorname{gph}(S^\nu)^{-1})$ and there exist $y^\nu \to \bar{y}$ and $x^\nu \to \bar{x}$ such that $x^\nu \in (S^\nu)^{-1}(y^\nu)$. We seek to establish that $\bar{x} \in \operatorname{LimInn}(S^\nu)^{-1}(\mathbb{B}(\bar{y},\varepsilon^\nu))$ for some vanishing ε^ν and need to construct $\bar{x}^\nu \to \bar{x}$, $\bar{y}^\nu \to \bar{y}$, and $\varepsilon^\nu \to 0$ such that $\bar{x}^\nu \in (S^\nu)^{-1}(\bar{y}^\nu)$ and $\bar{y}^\nu \in \mathbb{B}(\bar{y},\varepsilon^\nu)$. We see that $\bar{x}^\nu = x^\nu$, $\bar{y}^\nu = y^\nu$, and $\varepsilon^\nu = \|y^\nu - \bar{y}\|$ suffices. $\square$

The theorem establishes that cluster points of near-solutions of approximating generalized equations are indeed solutions of the actual generalized equation as long as the tolerance vanishes. Moreover, every such solution can be approached by near-solutions of the approximating generalized equations provided that the tolerance vanishes sufficiently slowly.

5.5 Example (homotopy methods for generalized equations). *For an osc set-valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ and $\bar{y} \in \mathbb{R}^n$, a homotopy method for solving the generalized equation $\bar{y} \in S(x)$ involves*

solving the approximating generalized equations

$$\bar{y}^\nu \in S^\nu(x) = (1-\lambda^\nu)S(x) + \lambda^\nu x$$

for $\lambda^\nu \in (0,1] \to 0$ *and* $\bar{y}^\nu \to \bar{y}$. *Trivially,* $\bar{y}^\nu$ *solves this equation when* $\lambda^\nu = 1$ *and this provides a starting point for iterations involving successively smaller* λ^ν. *The hope is that a gradual change in* λ^ν *results in approximating generalized equations that can be solve much quicker than the actual problem. In lieu of Theorem 5.4, we justify the method by establishing that* $S^\nu \xrightarrow{g} S$.

Detail. Let $(\hat{x},\hat{y}) \in \operatorname{LimOut}(\operatorname{gph} S^\nu)$. Then, there are $N \in \mathcal{N}_\infty^\#$, $x^\nu \xrightarrow[N]{} \hat{x}$, and $y^\nu \xrightarrow[N]{} \hat{y}$ with $y^\nu \in S^\nu(x^\nu)$. Consequently, $y^\nu = (1-\lambda^\nu)v^\nu + \lambda^\nu x^\nu$ for some $v^\nu \in S(x^\nu)$. Since $v^\nu = (1-\lambda^\nu)^{-1}(y^\nu - \lambda^\nu x^\nu) \xrightarrow[N]{} \hat{y}$ and S is osc, $\hat{y} \in S(\hat{x})$. This establishes that $\operatorname{LimOut}(\operatorname{gph} S^\nu) \subset \operatorname{gph} S$.

Next, take $(\hat{x},\hat{y}) \in \operatorname{gph} S$ and construct $x^\nu = \hat{x}$ and $y^\nu = (1-\lambda^\nu)\hat{y} + \lambda^\nu \hat{x}$. Then, $(x^\nu, y^\nu) \in \operatorname{gph} S^\nu$, $x^\nu \to \hat{x}$, and $y^\nu \to \hat{y}$, which means that $(\hat{x},\hat{y}) \in \operatorname{LimInn}(\operatorname{gph} S^\nu)$. This establishes that $\operatorname{LimInn}(\operatorname{gph} S^\nu) \supset \operatorname{gph} S$ and then also $S^\nu \xrightarrow{g} S$.

Figure 12 illustrates the approach when applied to $0 \in S(x) = \{x + \sin x + 1\}$, which is associated with the challenge that the derivative of $x \mapsto x + \sin x + 1$ vanishes for $x = \pi + 2\pi k$, $k \in \{\dots, -2, -1, 0, 1, 2, \dots\}$ and Newton's Method breaks down at these points. However, the approximation $S^\nu(x) = \{(1-\lambda^\nu)(x + \sin x + 1) + \lambda^\nu x\}$ doesn't have that problem for $\lambda^\nu > 0$. The figure shows the graph of S (solid line) and those of S^ν for $\lambda^\nu = 1, 0.5, 0.1$. □

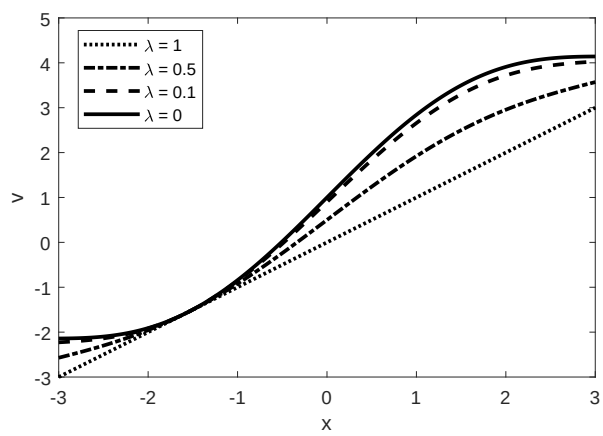

Figure 12: Graph of S (solid line) in Example 5.5 and those of its graphically converging approximations.

In the convex case, we have a useful connection due to Attouch [2] between epi-convergence of functions and graphical convergence of their sets of subgradients viewed as set-valued mappings.

5.6 Proposition (convergence for convex functions). *For proper, lsc, convex functions* $f^\nu, f : \mathbb{R}^n \to \overline{\mathbb{R}}$,

$$f^\nu \xrightarrow{e} f \iff \partial f^\nu \xrightarrow{g} \partial f \text{ and } f^\nu(x^\nu) \to f(\bar{x}) \text{ for some } v^\nu \to \bar{v}, x^\nu \to \bar{x} \text{ with } v^\nu \in \partial f^\nu(x^\nu) \text{ and } \bar{v} \in \partial f(\bar{x}).$$

6 Quantification of Set-Convergence

The rate of convergence of approximating solutions to an actual one relates to the rate with which the corresponding set-convergence takes place. This, in turn, requires us to clarify the meaning of distances between two sets. To allow for more flexibility, we define parallel to (1.1) the point-to-set distance $\operatorname{dist}(\bar{x}, C) = \inf_{x \in C} \|x - \bar{x}\|$ for a nonempty $C \subset \mathbb{R}^n$ using any norm $\|\cdot\|$ on $\mathbb{R}^n$ and $\operatorname{dist}(\bar{x}, \emptyset) = \infty$.

The *excess* of C over D is defined as

$$\operatorname{exs}(C; D) = \begin{cases} \sup_{x \in C} \operatorname{dist}(x, D) & \text{if } C \neq \emptyset, D \neq \emptyset \\ \infty & \text{if } C \neq \emptyset, D = \emptyset \\ 0 & \text{otherwise.} \end{cases}$$

Figure 13 shows that $\operatorname{exs}(C; D)$ can be different than $\operatorname{exs}(D; C)$ and it becomes natural to look at the sum of the two, as proposed by Pompeiu in 1905, or the larger of the two as promoted by Hausdorff in 1927. The latter has become known as the Hausdorff distance. In our setting with unbounded sets commonly occurring, either of these could be infinity. Thus, we adopt a truncation that can be traced back to Walkup and Wets [28] and Mosco [13] in their work on convex cones and convex sets, respectively. Its systematic study started with the work of Attouch and Wets [4].

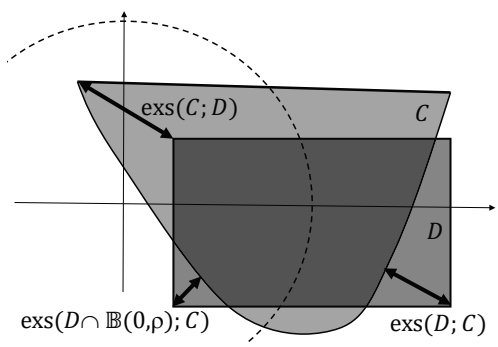

Figure 13: The excess of C over D and vice versa. The truncated Hausdorff distance $\hat{d\!l}_\rho(C, D) = \max\{\operatorname{exs}(C \cap \mathbb{B}(0, \rho); D), \ \operatorname{exs}(D \cap \mathbb{B}(0, \rho); C)\}$ considers only excess of the portion of a set inside a ball with radius ρ (dashed circle). Here, the Euclidean norm is assumed, but any other can be employed as well.

For any $\rho \geq 0$ and $C, D \subset \mathbb{R}^n$, the *truncated Hausdorff distance* between C and D is

$$\hat{d\!l}_\rho(C, D) = \max\Big\{ \operatorname{exs}\big(C \cap \mathbb{B}(0, \rho); D\big), \quad \operatorname{exs}\big(D \cap \mathbb{B}(0, \rho); C\big) \Big\}.$$

Figure 13 exemplifies the effect of truncation: $\operatorname{exs}(C \cap \mathbb{B}(0, \rho); D) = \operatorname{exs}(C; D)$ in this case because the portion of C furthest away from D is contained in $\mathbb{B}(0, \rho)$, while $\operatorname{exs}(D \cap \mathbb{B}(0, \rho); C)$ is quite a bit smaller than $\operatorname{exs}(D; C)$. Still, $\hat{d\!l}_\rho(C, D) = \operatorname{exs}(C; D)$ in this figure. Generally, $\hat{d\!l}_\rho(C, D)$ is finite for any nonempty C, D and $\rho \in \mathbb{R}_+$.

The truncated Hausdorff distance characterizes set-convergence: Given $\bar{\rho} \geq 0$ and closed $C \subset \mathbb{R}^n$,

$$C^\nu \xrightarrow{s} C \iff \hat{d\!l}_\rho(C^\nu, C) \to 0 \text{ for all } \rho \in [\bar{\rho}, \infty). \tag{6.1}$$

The need for considering all ρ sufficiently large is caused by the fact that the truncated Hausdorff distance would "miss" differences between C^ν and C if they only occur further away from the origin than ρ (in the norm used); see [18, Theorem 4.36] for a proof and details[14].

In particular, we are interested in distances between epigraphs as these relate to distances between minimizers and minimum values. For $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$, $\operatorname{epi} f$ and $\operatorname{epi} g$ are subsets of $\mathbb{R}^{n+1}$ and hence require a norm on $\mathbb{R}^{n+1}$. For any norm $\|\cdot\|$ on $\mathbb{R}^n$, a main choice for norm on $\mathbb{R}^{n+1}$ is

$$\max\big\{\|x-\bar{x}\|, |\alpha-\bar{\alpha}|\big\} \text{ for } (x,\alpha),(\bar{x},\bar{\alpha}) \in \mathbb{R}^n \times \mathbb{R}, \tag{6.2}$$

which typically results in the simplest formulae. In the context of minimization, we then obtain the following bounds; see [21, Proposition 2.2].

6.1 Theorem (approximation of minima and near-minimizers). *For $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$, $\varepsilon, \rho \in \mathbb{R}_+$, and any norm on $\mathbb{R}^n$ with* (6.2) *for $\mathbb{R}^{n+1}$, suppose that*

$$\inf f, \inf g \in [-\rho, \rho-\varepsilon), \quad \operatorname{argmin} f \cap \mathbb{B}(0,\rho) \neq \emptyset, \quad \operatorname{argmin} g \cap \mathbb{B}(0,\rho) \neq \emptyset.$$

Then,

$$\big|\inf f - \inf g\big| \leq \hat{d\!l}_\rho(\operatorname{epi} f, \operatorname{epi} g)$$
$$\operatorname{exs}\big(\varepsilon\text{-}\operatorname{argmin} g \cap \mathbb{B}(0,\rho);\ \delta\text{-}\operatorname{argmin} f\big) \leq \hat{d\!l}_\rho(\operatorname{epi} f, \operatorname{epi} g) \text{ for } \delta > \varepsilon + 2\hat{d\!l}_\rho(\operatorname{epi} f, \operatorname{epi} g).$$

The theorem holds for nearly arbitrary functions as long as minimizers are attained with finite minimum values and ρ is large enough. The bounds are sharp as discussed in [19, 21].

Computation of the truncated Hausdorff distance is supported by the following condition that originated with Kenmochi [10] and fully realized by Attouch and Wets [4]. For the present form, see [21, Proposition 4.1]. We denote *lower level-sets* of $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ by $\{f \leq \delta\} = \{x \in \mathbb{R}^n \mid f(x) \leq \delta\}$ for any $\delta \in \overline{\mathbb{R}}$.

6.2 Proposition (Kenmochi condition). *For $f, g : \mathbb{R}^n \to \overline{\mathbb{R}}$, both with nonempty epigraphs, $\rho \in \mathbb{R}_+$, and any norm on $\mathbb{R}^n$ with* (6.2) *for $\mathbb{R}^{n+1}$, we have*

$$\hat{d\!l}_\rho(\operatorname{epi} f, \operatorname{epi} g) = \inf\Big\{\eta \geq 0 \;\Big|\; \inf_{x\in\mathbb{B}(\bar{x},\eta)} g(x) \leq \max\{f(\bar{x}), -\rho\} + \eta, \quad \forall \bar{x} \in \{f \leq \rho\} \cap \mathbb{B}(0,\rho)$$
$$\inf_{x\in\mathbb{B}(\bar{x},\eta)} f(x) \leq \max\{g(\bar{x}), -\rho\} + \eta, \quad \forall \bar{x} \in \{g \leq \rho\} \cap \mathbb{B}(0,\rho)\Big\}.$$

From this fact, we can develop several useful bounds on the truncated Hausdorff distance. We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *Lipschitz continuous* with modulus $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ if

$$\big|f(x) - f(\bar{x})\big| \leq \kappa(\rho)\|x-\bar{x}\|_2 \text{ for all } \rho \in \mathbb{R}_+ \text{ and } x, \bar{x} \in \mathbb{B}_2(0,\rho) = \big\{x \in \mathbb{R}^n \;\big|\; \|x\|_2 \leq \rho\big\}.$$

Note that the modulus is permitted to depend on the size of the ball on which the condition holds.

[14] Any norm can be used to underpin the right-hand side of (6.1), but this is also the case for the left-hand side. Even though Definition 1.1 is stated in terms of $\|\cdot\|_2$, any norm could have been adopted because, by Proposition 2.1, it's only the topology on $\mathbb{R}^n$ that matters and all norms produce the same topology in finite dimensions.

6.3 Proposition (distance between composite functions). *For proper $f_0, g_0 : \mathbb{R}^n \to \overline{\mathbb{R}}$, consider $f = f_0 + h \circ F$ and $g = g_0 + h \circ G$, where $F = (f_1, \ldots, f_m)$, $G = (g_1, \ldots, g_m)$,*

(a) *$h : \mathbb{R}^m \to \mathbb{R}$ is Lipschitz continuous with modulus $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$, and*

(b) *$f_i, g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, are Lipschitz continuous with common modulus $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$.*

If $\|\cdot\|_2$ is the norm on $\mathbb{R}^n$ and (6.2) with $\|\cdot\| = \|\cdot\|_2$ is the norm on $\mathbb{R}^{n+1}$, then for any $\rho \in \mathbb{R}_+$,

$$\hat{d}_\rho(\operatorname{epi} f, \operatorname{epi} g) \le \big(1 + \sqrt{m}\kappa(\rho^*)\lambda(\hat{\rho})\big)\hat{d}_{\bar{\rho}}(f_0, g_0) + \kappa(\rho^*) \sup_{x \in \mathbb{B}_2(0,\rho)} \big\|F(x) - G(x)\big\|_2,$$

where $\bar{\rho} \ge \rho + \max\{\sup_{x \in \mathbb{B}_2(0,\rho)} |h(F(x))|, \sup_{x \in \mathbb{B}_2(0,\rho)} |h(G(x))|\}$, $\hat{\rho} > \rho + \hat{d}_{\bar{\rho}}(\operatorname{epi} f_0, \operatorname{epi} g_0)$, and $\rho^* \ge \max\{\sup_{x \in \mathbb{B}_2(0,\hat{\rho})} \|F(x)\|_2, \sup_{x \in \mathbb{B}_2(0,\hat{\rho})} \|G(x)\|_2\}$.

Proof. Our goal is to establish the two conditions on the right-hand side in Proposition 6.2. Let $\mathbb{S}(\bar{\rho}) = \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid \|x\|_2 \le \bar{\rho}, |\alpha| \le \bar{\rho}\}$ be a centered ball using the assumed norm on $\mathbb{R}^{n+1}$. Set $\eta = \hat{d}_{\bar{\rho}}(\operatorname{epi} f_0, \operatorname{epi} g_0)$, $\varepsilon \in (0, \hat{\rho} - \rho - \eta)$, and $\bar{x} \in \{f \le \rho\} \cap \mathbb{B}_2(0, \rho)$. Then, $f_0(\bar{x}) \le \rho - h(F(\bar{x})) \le \bar{\rho}$.

First, suppose that $f_0(\bar{x}) \ge -\bar{\rho}$ so that $(\bar{x}, f_0(\bar{x})) \in \operatorname{epi} f_0 \cap \mathbb{S}(\bar{\rho})$. Consequently, there exists $(\hat{x}, \hat{\alpha}) \in \operatorname{epi} g_0$ with $\|\bar{x} - \hat{x}\|_2 \le \eta + \varepsilon$ and $|\hat{\alpha} - f_0(\bar{x})| \le \eta + \varepsilon$. Thus, $g_0(\hat{x}) \le \hat{\alpha} \le f_0(\bar{x}) + \eta + \varepsilon$ and

$$\begin{aligned}
&\inf_{x \in \mathbb{B}(\bar{x}, \eta+\varepsilon)} g(x) \le g_0(\hat{x}) + h\big(G(\hat{x})\big) = g_0(\hat{x}) + h\big(F(\bar{x})\big) + h\big(G(\hat{x})\big) - h\big(G(\bar{x})\big) + h\big(G(\bar{x})\big) - h\big(F(\bar{x})\big) \\
&\le f_0(\bar{x}) + \eta + \varepsilon + h\big(F(\bar{x})\big) + \kappa(\rho^*)\big\|G(\hat{x}) - G(\bar{x})\big\|_2 + \kappa(\rho^*)\big\|F(\bar{x}) - G(\bar{x})\big\|_2 \\
&\le f_0(\bar{x}) + h\big(F(\bar{x})\big) + \eta + \varepsilon + \kappa(\rho^*)\sqrt{m} \max_{i=1,\ldots,m} \big|g_i(\hat{x}) - g_i(\bar{x})\big| + \kappa(\rho^*) \sup_{x \in \mathbb{B}_2(0,\rho)} \big\|F(x) - G(x)\big\|_2 \\
&\le f_0(\bar{x}) + h\big(F(\bar{x})\big) + \eta + \varepsilon + \kappa(\rho^*)\sqrt{m}\lambda(\hat{\rho})(\eta + \varepsilon) + \kappa(\rho^*) \sup_{x \in \mathbb{B}_2(0,\rho)} \big\|F(x) - G(x)\big\|_2 \\
&\le \max\big\{f_0(\bar{x}) + h\big(F(\bar{x})\big), -\rho\big\} + \big(1 + \kappa(\rho^*)\sqrt{m}\lambda(\hat{\rho})\big)(\eta + \varepsilon) + \kappa(\rho^*) \sup_{x \in \mathbb{B}_2(0,\rho)} \big\|F(x) - G(x)\big\|_2.
\end{aligned}$$

Second, suppose that $f_0(\bar{x}) < -\bar{\rho}$. Then, $(\bar{x}, -\bar{\rho}) \in \operatorname{epi} f_0 \cap \mathbb{S}(\bar{\rho})$ and there exists $(\hat{x}, \hat{\alpha}) \in \operatorname{epi} g_0$ with $\|\bar{x} - \hat{x}\|_2 \le \eta + \varepsilon$ and $|\hat{\alpha} + \bar{\rho}| \le \eta + \varepsilon$. Thus, $g_0(\hat{x}) \le \hat{\alpha} \le -\bar{\rho} + \eta + \varepsilon$ and, similar to above,

$$\begin{aligned}
\inf_{x \in \mathbb{B}(\bar{x}, \eta+\varepsilon)} g(x) &\le -\bar{\rho} + h\big(F(\bar{x})\big) + \eta + \varepsilon + \kappa(\rho^*)\sqrt{m}\lambda(\hat{\rho})(\eta + \varepsilon) + \kappa(\rho^*) \sup_{x \in \mathbb{B}_2(0,\rho)} \big\|F(x) - G(x)\big\|_2 \\
&\le \max\big\{f_0(\bar{x}) + h\big(F(\bar{x})\big), -\rho\big\} + \big(1 + \kappa(\rho^*)\sqrt{m}\lambda(\hat{\rho})\big)(\eta + \varepsilon) + \kappa(\rho^*) \sup_{x \in \mathbb{B}_2(0,\rho)} \big\|F(x) - G(x)\big\|_2.
\end{aligned}$$

The second inequality follows because $-\bar{\rho} + h(F(\bar{x})) \le -\bar{\rho} + \sup_{x \in \mathbb{B}_2(0,\rho)} |h(F(x))| \le -\rho$. Thus, in both cases, we obtain the same upper bound on $\inf_{x \in \mathbb{B}(\bar{x}, \eta+\varepsilon)} g(x)$. Repeating these arguments with the roles of f and g switched, we obtain the result via Proposition 6.2 after letting ε tend to zero. ◻

For other rules to compute the truncated Hausdorff distance, we refer to [6, 4, 18, 19, 21].

We next turn to solution errors for generalized equations. If $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, then $\operatorname{gph} S \subset \mathbb{R}^n \times \mathbb{R}^m$ and the truncated Hausdorff distance between such sets requires a norm on $\mathbb{R}^n \times \mathbb{R}^m$. Given any norm $\|\cdot\|_a$ on $\mathbb{R}^n$ and any norm $\|\cdot\|_b$ on $\mathbb{R}^m$, a main choice is

$$\max\big\{\|x\|_a, \|y\|_b\big\} \text{ for } (x, y) \in \mathbb{R}^n \times \mathbb{R}^m. \tag{6.3}$$

This leads to the following bound on near-solutions proven in [21, Theorem 5.1].

6.4 Theorem (near-solutions of generalized equations). *For $S, T : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, with nonempty graphs, and any norms on $\mathbb{R}^n$ and $\mathbb{R}^m$ with (6.3) for $\mathbb{R}^n \times \mathbb{R}^m$, suppose that $0 \le \varepsilon \le \rho < \infty$ and $\bar{y} \in \mathbb{B}(0, \rho - \varepsilon)$. If $\delta > \varepsilon + \hat{\mathbb{d}}_\rho(\operatorname{gph} S, \operatorname{gph} T)$, then*

$$\operatorname{exs}\Big(S^{-1}\big(\mathbb{B}(\bar{y}, \varepsilon)\big) \cap \mathbb{B}(0, \rho);\ T^{-1}\big(\mathbb{B}(\bar{y}, \delta)\big)\Big) \le \hat{\mathbb{d}}_\rho(\operatorname{gph} S, \operatorname{gph} T).$$

If $\operatorname{gph} T$ is closed, then the result also holds with $\delta = \varepsilon + \hat{\mathbb{d}}_\rho(\operatorname{gph} S, \operatorname{gph} T)$.

The theorem is sharp. For example, consider $S, T : \mathbb{R} \rightrightarrows \mathbb{R}$ with $S(x) = [x, \infty)$ when $x \in [0, 1]$ and $S(x) = \emptyset$ otherwise; and $T(x) = (1, \infty)$ when $x \in [1, 2]$ and $T(x) = \emptyset$ otherwise. Then for $\rho \ge 0$, $\hat{\mathbb{d}}_\rho(\operatorname{gph} S, \operatorname{gph} T) = 1$, $S^{-1}(0) = \{0\}$, $T^{-1}(\delta) = [1, 2]$, and $\operatorname{exs}(S^{-1}(0) \cap \mathbb{B}(0, \rho); T^{-1}(\mathbb{B}(0, \delta)) = 1$ when $\delta > 1$. When $\delta \le 1$, the excess becomes infinity because $T^{-1}(\delta) = \emptyset$. If T is modified to having $T(x) = [1, \infty)$ for $x \in [1, 2]$, then $\delta = 1$ gives an excess of one.

6.5 Example (stationarity of composite functions). *For a proper lsc function $\varphi : \mathbb{R}^m \to \overline{\mathbb{R}}$ and a smooth mapping $F : \mathbb{R}^n \to \mathbb{R}^m$, the composite function $\varphi \circ F$ has $0 \in \nabla F(x)^\top \partial\varphi(F(x))$ as a necessary optimality condition under standard assumptions [18, Theorem 10.6], where the $m \times n$-matrix $\nabla F(x)$ is the Jacobian of F at x. Instead of solving this generalized equation, we may have to settle for an approximation. The error caused by this switch can be quantified using Theorem 6.4.*

Detail. By introducing auxiliary vectors $y, z \in \mathbb{R}^m$, the optimality condition is equivalently stated in terms of the set-valued mapping $S : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m \rightrightarrows \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^n$ as $0 \in S(x, y, z)$, with

$$S(x, y, z) = \big\{F(x) - z\big\} \times \big\{\partial\varphi(z) - y\big\} \times \big\{\nabla F(x)^\top y\big\}.$$

Since $0 \in S(x, y, z)$ is also an optimality condition for the problem of minimizing $\varphi(z)$ subject to $F(x) = z$, we see that y can be interpreted as a multiplier vector and z as representing feasibility.

An approximating function $\psi \circ G$ expressed in terms of proper lsc $\psi : \mathbb{R}^m \to \overline{\mathbb{R}}$ and smooth $G : \mathbb{R}^n \to \mathbb{R}^m$ has parallel optimality conditions: $0 \in T(x, y, z)$, where

$$T(x, y, z) = \big\{G(x) - z\big\} \times \big\{\partial\psi(z) - y\big\} \times \big\{\nabla G(x)^\top y\big\}.$$

In view of Theorem 6.4, a bound on $\hat{\mathbb{d}}_\rho(\operatorname{gph} S, \operatorname{gph} T)$ leads to an estimate of the change in near-stationary points as we pass from $\varphi \circ F$ to $\psi \circ G$.

The "input space" for S, T is $\mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ on which we adopt the norm $\max\{\|x\|_2, \|y\|_2, \|z\|_2\}$ for $(x, y, z) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$. The "output space" $\mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^n$ is assigned the norm $\max\{\|u\|_2, \|v\|_2, \|w\|_2\}$ for $(u, v, w) \in \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^n$. Then, by [21, Theorem 5.3], for $\rho \in \mathbb{R}_+$,

$$\hat{\mathbb{d}}_\rho(\operatorname{gph} S, \operatorname{gph} T) \le \sup_{\|x\|_2 \le \rho} \max\Big\{\big\|G(x) - F(x)\big\|_2 + \hat{\mathbb{d}}_{2\rho}\big(\operatorname{gph} \partial\varphi, \operatorname{gph} \partial\psi\big),\ \rho\big\|\nabla G(x) - \nabla F(x)\big\|_F\Big\},$$

where $\|\cdot\|_F$ is the Frobenius norm. In calculating distances between $\operatorname{gph} \partial\varphi$ and $\operatorname{gph} \partial\psi$, we adopt the norm $\max\{\|z\|_2, \|y\|_2\}$ for $(z, y) \in \mathbb{R}^m \times \mathbb{R}^m$. $\square$

7 Extensions

The quantification of set-convergence hints to the possibility of constructing a metric on spaces of sets and thereby enabling general metric space theory. Although $\hat{d}_\rho$ isn't a metric because it fails the triangle inequality and there is also the issue with discrepancies beyond $\mathbb{B}(0,\rho)$, the closely related Attouch-Wets distance[15] metrizes set-convergence on the space of nonempty closed sets. The resulting metric space is complete, separable, and finitely compact. The latter means that every closed ball in that space is compact. This points to the following fact available directly from the properties of set-convergence: For every sequence $\{C^\nu \subset \mathbb{R}^n, n \in \mathbb{N}\}$, either $C^\nu \xrightarrow{s} \emptyset$ or there exists $N \in \mathcal{N}_\infty^\#$ and a nonempty set C such that $\{C^\nu, \nu \in N\} \xrightarrow{s} C$. This result can be traced back to Zoretti in 1909, a student of Painlevé, but here presented in the form found in [18, Theorem 4.18].

When restricted to epigraphs of lsc functions with nonempty domains, the Attouch-Wets distance furnishes a metric on the space of such functions with the property that convergence of functions implies convergence of the corresponding minimizers and minimum values in the sense of Theorem 3.7(a,b,c). This can be utilized to analyze infinite-dimensional problems in nonparametric statistics and elsewhere; see [22, 25, 20]. The compactness of closed and bounded subsets of such spaces trivializes questions of existence of solutions, results in finite covering numbers [20], and facilitates the construction of approximating functions that are computationally attractive [22, 19, 20].

The extension of set-convergence and its quantification from subsets of $\mathbb{R}^n$ to those of general metric spaces is trivial as the definitions only require a notion of distance between points. The truncated Hausdorff distance is then defined relative to an arbitrary center point; for $\mathbb{R}^n$ we simply adopt the origin, which gives rise to $\mathbb{B}(0,\rho)$ in the definition. In fact, even on $\mathbb{R}^n$, it might be beneficial to move the center point elsewhere to better capture the scale of a problem; see [19, 20]. A complication beyond $\mathbb{R}^n$ is that (6.1) may hold only with $\Longleftarrow$; see [19] for an introduction and [7] for details. Nevertheless, set-convergence remains an important tool in the infinite-dimensional setting; see for example [5, 15, 11, 3, 14, 19].

Acknowledgement. This work is supported by ONR (Operations Research) under N0001420WX00519 and AFOSR (Optimization and Discrete Mathematics) under F4FGA08272G001.

References

[1] A. Aswani. Statistics with set-valued functions: Applications to inverse approximate optimization. *Mathematical Programming B*, 174(1-2):225–251, 2019.

[2] H. Attouch. Convergence de fonctions convexes, des sous-différentiels et semi-groupes associés. *Comptes Rendus de l'Académie des Sciences de Paris*, 284(10):539–542, 1977.

[15] This distance emerged from the work of Attouch and Wets in the early 1990s; see [18, Section 4.I] where it's called the integrated set distance.

[3] H. Attouch, G. Buttazzo, and G. Michaille. *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*. MOS-SIAM Series on Optimization. SIAM, 2nd edition, 2014.

[4] H. Attouch and R. J-B Wets. Quantitative stability of variational systems: I. The epigraphical distance. *Transactions of the American Mathematical Society*, 328(2):695–729, 1991.

[5] H. Attouch and R. J-B Wets. Quantitative stability of variational systems: II. A framework for nonlinear conditioning. *SIAM J. Optimization*, 3:359–381, 1993.

[6] D. Aze and J.-P. Penot. Operations on convergent families of sets and functions. *Optimization*, 21(4):521–534, 1990.

[7] G. Beer. *Topologies on Closed and Closed Convex Sets*, volume 268 of *Mathematics and its Applications*. Kluwer, 1992.

[8] E. De Giorgi. Γ-convergenza e G-convergenza. *Bollettino dell' Unione Matematica Italiana*, 14-A:213–220, 1977.

[9] E. De Giorgi and T. Franzoni. Su un tipo di convergenza variazionale. *Atti Accademia Nazionale de Lincei. Rendiconti della Classe di Scienze Fisiche, Matematiche e Naturali*, 58:842–850, 1975.

[10] N. Kenmochi. The semi-discretization method and time dependent parabolic variational inequalities. *Proceedings of the Japan Academy, Series A, Mathematical Sciences*, 50(9):714–717., 1974.

[11] L. A. Korf and R. J-B Wets. Random lsc functions: an ergodic theorem. *Mathematics of Operations Research*, 26(2):421–445, 2001.

[12] J. J. Moreau. Approximation en graphe d'une évolution discontinue. *RAIRO*, 12(1):75–84, 1978.

[13] U. Mosco. Convergence of convex sets and of solutions of variational inequalities. *Advances in Mathematics*, 3:510–585, 1969.

[14] C. Phelps, J. O. Royset, and Q. Gong. Optimal control of uncertain systems using sample average approximations. *SIAM J. Control and Optimization*, 54(1):1–29, 2016.

[15] E. Polak. *Optimization. Algorithms and Consistent Approximations*, volume 124 of *Applied Mathematical Sciences*. Springer, 1997.

[16] S. M. Robinson. Normal maps induced by linear transformations. *Mathematics of Operations Research*, 17(3):691–714, 1992.

[17] R. T. Rockafellar and J. O. Royset. Random variables, monotone relations, and convex analysis. *Mathematical Programming B*, 148(1):297–331, 2014.

[18] R.T. Rockafellar and R. J-B Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaft*. Springer, 3rd printing-2009 edition, 1998.

[19] J. O. Royset. Approximations and solution estimates in optimization. *Mathematical Programming*, 170(2):479–506, 2018.

[20] J. O. Royset. Approximations of semicontinuous functions with applications to stochastic optimization and statistical estimation. *Mathematical Programming*, OnlineFirst, 2020.

[21] J. O. Royset. Stability and error analysis for optimization and generalized equations. *SIAM J. Optimization*, OnlineFirst, 2020.

[22] J. O. Royset and R. J-B Wets. Multivariate epi-splines and evolving function identification problems. *Set-Valued and Variational Analysis*, 24(4):517–545, 2016. Erratum: pp. 547-549.

[23] J. O. Royset and R. J-B Wets. Variational theory for optimization under stochastic ambiguity. *SIAM J. Optimization*, 27(2):1118–1149, 2017.

[24] J. O. Royset and R. J-B Wets. Lopsided convergence: an extension and its quantification. *Mathematical Programming*, 177(1):395–423, 2019.

[25] J. O. Royset and R. J-B Wets. Variational analysis of constrained M-estimators. *Annals of Statistics*, to appear, 2020.

[26] S. Saha and A. Guntuboyina. On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *Annals of Statistics*, to appear, 2020.

[27] S. Spagnolo. Convergence in energy for elliptic operators. In *Proceedings of the Third Symposium on the Numerical Solutions of Partial Differential Equations*, pages 469–498. Academic Press, 1976.

[28] D. W. Walkup and R. J-B Wets. Continuity of some convex-cone-valued mappings. *Proceedings of the American Mathematical Society*, 18:229–235, 1967.

[29] R. J-B Wets. Convergence of convex functions, variational inequalities, and convex optimization problems. In R. Cottle, R. Giannessi, and J.-L. Lions, editors, *Varitional Inequalities and Complementary Problems*, pages 375–403. Wiley, New York, 1980.

[30] R. A. Wijsman. Convergence of sequences of convex sets, cones and functions. *Bulletin of the American Mathematical Society*, 70:186–188, 1964.

[31] R. A. Wijsman. Convergence of sequences of convex sets, cones and functions. II. *Transactions of the American Mathematical Society*, 123:32–45, 1966.