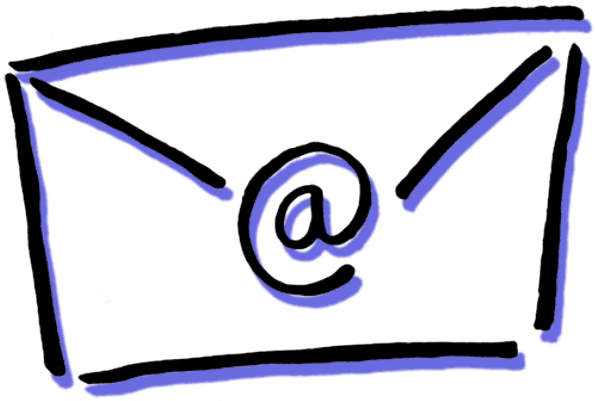
SIAG/OPT Views and News

A Forum for the SIAM Activity Group on Optimization

Volume 28 Number 1 December 2020

Contents

Are you receiving this by postal mail?
Do you prefer electronic delivery?



Email siagoptnews@lists.mcs.anl.gov!

Articles

Subgradient methods under weak convexity and tame geometry

Damek Davis
School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850
people.orie.cornell.edu/dsd95

Dmitriy Drusvyatskiy
Department of Mathematics, University of Washington, Seattle, WA 98195
www.math.washington.edu/~ddrusv

1 Introduction

The subgradient method is a classical algorithm for minimizing a nonsmooth Lipschitz continuous function φ on $\mathbb{R}^d$. Starting from an initial iterate x_0, the method simply iterates

$$x_{t+1} = x_t - \alpha_t v_t \qquad \text{with } v_t \in \partial\varphi(x_t). \tag{1.1}$$

Here, the positive sequence $\{\alpha_t\}_{t\geq 0}$ is user specified and the set $\partial\varphi(x)$ is the *Clarke subdifferential*, which is defined as

$$\partial\varphi(x) = \operatorname{conv}\left\{\lim_{i\to\infty} \nabla\varphi(x_i) : x_i \to x,\ x_i \in \operatorname{dom}(\nabla\varphi)\right\}.$$

In classical circumstances, the subdifferential reduces to more familiar objects. For example, when φ is C^1-smooth at x, the subdifferential $\partial\varphi(x)$ consists only of the gradient $\nabla\varphi(x)$, while for convex functions, it reduces to the subdifferential in the sense of convex analysis. It is precisely under these two settings—smoothness and/or convexity—that the subgradient method is most well-understood [32, 33, 42, 43].

While smooth and convex problems encompass a variety of applications, problems lacking both qualities have recently emerged in modern machine learning practice. Indeed, industry-backed solvers, such as Google's TensorFlow and Facebook's PyTorch, now routinely train nonsmooth deep networks via (stochastic) subgradient methods, powering widespread empirical success. It is important to note, however, that the subgradient method on general Lipschitz

continuous functions may fail to find any critical point, due to highly pathological examples which would never appear in practice [19]. Consequently to make progress, it is essential to restrict the problem class under consideration.

Inspired by the success of the subgradient method in applications, several recent works have revisited the foundations of subgradient methods, identifying two amenable problem classes: *weakly convex* and *tame* problems. The weakly convex class is broad, capturing important tasks in (robust) statistical estimation. As we will survey in this article, the subgradient method on weakly convex problems admits strong iteration complexity guarantees. Though broad, the weakly convex class does not capture modern nonsmooth neural networks. For this we turn to tame functions, a virtually exhaustive class of non pathological functions, including all semi-algebraic and semi-analytic functions. Though considerably broader than the weakly convex class, available complexity guarantees for tame functions are much weaker, and instead existing work illuminates only asymptotic behavior.

The purpose of this article is to survey the subgradient method for weakly convex and tame problems, illustrating two elegant tools for analyzing subgradient dynamics:

1. The Moreau envelope as a Lyapunov function for weakly convex problems.

2. The chain rule for tame functions and its role in guaranteeing convergence.

The first tool is elementary and has found wide applicability. The second tool is based on a more nuanced machinery, rooted in tame geometry [53, 55, 56], and applies more broadly.

2 Weak Convexity and the Moreau Envelope

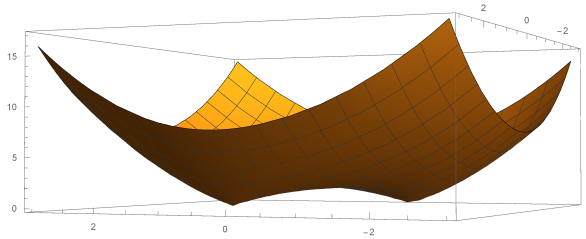
We begin with the following seemingly simple question:

How can one judge the performance of the subgradient method for nonsmooth and nonconvex problems?

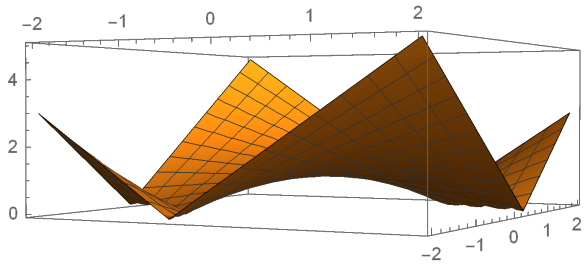
The reason this question is nontrivial is that the classical literature exhibits a dichotomy between convex and smooth settings. Namely, convex optimization algorithms are judged by the rate at which they decrease the function value along the iterates, while smooth optimization algorithms are instead judged by the magnitude of the gradients. Neither performance metric is appropriate for analyzing the subgradient method on functions that are *simultaneously* nonsmooth and nonconvex. Indeed, since the problem is nonconvex, the functional suboptimality gap, $\varphi(x_t) - \inf \varphi$ need not tend to zero. Moreover, the nonsmooth stationarity measure, $\operatorname{dist}(0; \partial\varphi(x_t))$ may remain bounded away from zero along the iterates even for convex problems, as the example $\varphi(x) = |x|$ illustrates. While salvaging the first measure is hopeless, the second measure can in some cases be salvaged if we are willing to perturb slightly the point x_t at which $\partial\varphi$ is evaluated. The message of this section is that for weakly convex problems, this may be done in a principled way through *implicit smoothing*.

2.1 The weakly convex class

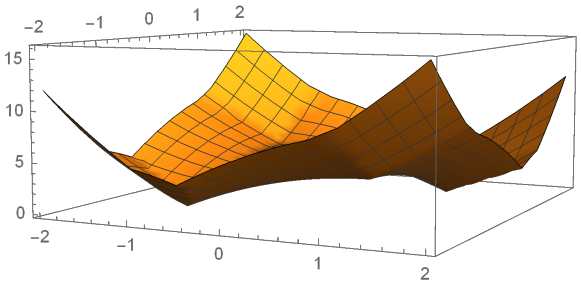
A function φ is called *ρ-weakly convex* if the perturbed function $x \mapsto \varphi(x) + \frac{\rho}{2}\|x\|^2$ is convex.[1] This function class is broad and includes convex functions, smooth functions with Lipschitz continuous gradient, and any function of the form $\varphi = h \circ c$, with h convex and Lipschitz and c a smooth map with Lipschitz Jacobian. Classical literature highlights the importance of weak convexity in optimization [48, 49, 51], while recent advances in statistical learning and signal processing have further reinvigorated the problem class. For example, nonlinear least squares, phase retrieval [22, 28, 29], graph synchronization [1, 6, 54], and robust principal component analysis [15, 16] naturally lead to weakly convex formulations—see Figure 1. We refer the reader to the papers [4, 17, 20] and the previous SIAM news article [24] for detailed examples.



(a) $\varphi(x) = \mathbb{E}|(a^\top x)^2 - (a^\top \mathbf{1})^2|$ (phase retrieval)



(b) $\varphi(x, y) = \mathbb{E}|axyb - ab|$ (blind deconvolution)



(c) $\varphi(x) = \|xx^\top - \mathbf{1}\mathbf{1}^\top\|_1$ (robust PCA)

Figure 1: Examples of weakly convex functions and their use in practice.

The class of weakly convex problem is amenable to analysis precisely because it entails a simplification of the subdifferential. Specifically, ρ-weak convexity automatically guarantees that subgradients provide quadratic under-estimators with

[1] Weakly convex functions also go by other names such as lower-C^2, uniformly prox-regular, t^2-paraconvex, and semiconvex.

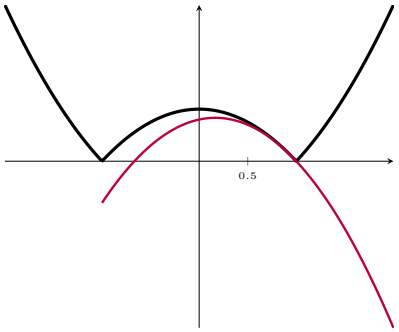


Figure 2: Quadratic under-estimator of $\varphi(x) = |x^2 - 1|$.

uniform amplitude (see Figure 2):

$$\varphi(y) \geq \varphi(x) + \langle v, y - x\rangle - \frac{\rho}{2}\|y - x\|^2, \tag{2.1}$$

for all $x, y \in \mathbb{R}^d$ and $v \in \partial\varphi(x)$. This fact is crucial for analyzing Algorithm (1.1), as we now discuss.

2.2 Complexity of Algorithm (1.1) for weakly convex problems

It has been known since Nurminskii's seminal work [46, 47] that when φ is ρ-weakly convex, the subgradient method generates an iterate sequence that subsequentially converges to a stationary point of the problem. Nonetheless, the complexity of the basic method and of its proximal extension has remained elusive until the recent work [20]. What appeared to be missing from prior work was a continuous measure of stationarity to monitor, instead of the highly discontinuous function $x \mapsto \mathrm{dist}(0; \partial\varphi(x))$. The strategy proposed in [20] relies on the elementary observation: weakly convex problems naturally admit a continuous measure of stationarity through implicit smoothing.

Setting the stage, for any $\lambda > 0$ define the *Moreau envelope* and the *proximal map*:

$$\varphi_\lambda(x) := \min_y \; \left\{\varphi(y) + \tfrac{1}{2\lambda}\|y - x\|^2\right\},$$
$$\mathrm{prox}_{\lambda\varphi}(x) := \arg\min_y \; \left\{\varphi(y) + \frac{1}{2\lambda}\|y - x\|^2\right\}.$$

Standard results [41] show that as long as $\lambda < \rho^{-1}$, the envelope φ_λ is C^1-smooth with the gradient given by

$$\nabla\varphi_\lambda(x) = \lambda^{-1}(x - \mathrm{prox}_{\lambda\varphi}(x)). \tag{2.2}$$

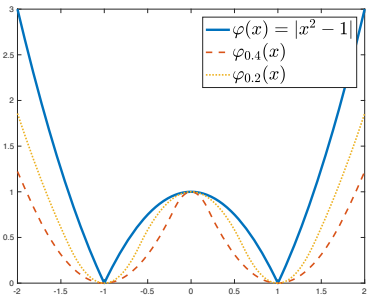
See Figure 3 for an illustration. This gradient $\nabla\varphi_\lambda(x)$ is closely related to the subdifferential of φ itself. Indeed, when φ is smooth, the norm $\|\nabla\varphi_\lambda(x)\|$ is proportional to the magnitude of the true gradient $\|\nabla\varphi(x)\|$. In the broader nonsmooth setting, the norm of the gradient $\|\nabla\varphi_\lambda(x)\|$ has an intuitive interpretation in terms of near-stationarity for the target problem. Namely, the definition of the Moreau envelope directly implies that for any point $x \in \mathbb{R}^d$, the proximal point $\hat{x} := \mathrm{prox}_{\lambda\varphi}(x)$ satisfies

$$\begin{cases} \|\hat{x} - x\| & = \lambda\|\nabla\varphi_\lambda(x)\|, \\ \varphi(\hat{x}) & \leq \varphi(x), \\ \mathrm{dist}(0; \partial\varphi(\hat{x})) & \leq \|\nabla\varphi_\lambda(x)\|. \end{cases} \tag{2.3}$$

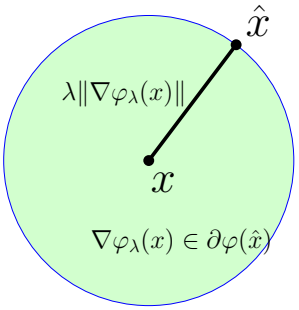
In summary,

> a small gradient $\|\nabla\varphi_\lambda(x)\|$ implies that x is *near* some point $\hat{x}$ that is *nearly stationary* for (2.6).

Moreover, for any point x, one can in principle estimate $\hat{x}$ to high precision by solving the *strongly convex* optimization problem defining the proximal operator. For a longer discussion of the near-stationarity concept, see [24] or [26, §4.1].



(a) Moreau envelope of $\varphi(x) = |x^2 - 1|$



(b) Approximate stationarity

Figure 3: An illustration of the Moreau envelope

The Moreau envelope not only provides a natural continuous measure of stationarity for weakly convex problems, it also serves as an approximate Lyapunov function for the subgradient dynamics. Indeed, the key observation of [20] is that Algorithm (1.1) can be interpreted as an approximate descent method on the Moreau envelope:

$$\varphi_\lambda(x_{t+1}) \leq \varphi_\lambda(x_t) - \alpha_t c_1\|\nabla\varphi_\lambda(x_t)\|^2 + \alpha_t^2 c_2, \tag{2.4}$$

where c_1, c_2 are problem dependent constants and $\lambda < \rho^{-1}$ is an arbitrary parameter. Rearranging and summing immediately yields a convergence rate:

Corollary 2.1. *The subgradient method will find a point x satisfying $\|\nabla\varphi_\lambda(x)\| \leq \epsilon$ using $O(\epsilon^{-4})$ subgradient evaluations.*

Thus, the gradient of the Moreau envelope tends to zero at a controlled rate along the sequence produced by (1.1), even

though the stationarity measure $\operatorname{dist}(0, \partial\varphi(x_t))$ may remain bounded below for all t. At first glance, this is somewhat surprising as neither the Moreau envelope nor the proximal operator appear in Algorithm (1.1).

An intuitive explanation of why (2.4) holds is that the subgradients of φ and the gradient $\nabla\varphi_\lambda$ are well-aligned:

$$\langle v, \nabla\varphi_\lambda(x)\rangle \geq \frac{1-\lambda\rho}{2}\|\nabla\varphi_\lambda(x)\|^2 \quad \text{for all } v \in \partial\varphi(x). \quad (2.5)$$

This estimate follows quickly: setting $\hat{x} := \operatorname{prox}_{\lambda\varphi}(x)$ and $y := \hat{x}$ in (2.1) yields

$$\begin{aligned}\varphi(\hat{x}) &\geq \varphi(x) + \langle v, \hat{x} - x\rangle - \frac{\rho}{2}\|\hat{x} - x\|^2 \\ &\geq \varphi(\hat{x}) + \langle v, \hat{x} - x\rangle + \frac{\lambda^{-1}-\rho}{2}\|\hat{x} - x\|^2.\end{aligned}$$

Recalling the relationship (2.2) immediately yields (2.5).

2.3 Extensions

Stochastic Model-Based Methods. The highlighted ideas generalize to a wide class of algorithms for stochastic and regularized problems:

$$\min_x \ \varphi(x) := f(x) + r(x) \quad \text{where} \quad f(x) = \mathbb{E}_{z\sim P}[f(x,z)]. \quad (2.6)$$

Here, z encodes the population data, which is assumed to follow some fixed but unknown probability distribution P. The functions f and r play qualitatively different roles. Typically, $f(x,z)$ evaluates the loss of the decision rule parametrized by x on a data point z. In contrast, the function $r\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ models constraints on the parameters x or encourages x to have some low dimensional structure, such as sparsity or low rank.

The article [20] analyzes generic algorithms that in each iteration t draw a sample $z_t \sim P$, approximate the loss $f(x, z_t)$ with a local model $f_{x_t}(y, z_t)$, and perform the update:

$$x_{t+1} = \arg\min_y \left\{ f_{x_t}(y, z_t) + r(y) + \frac{1}{2\alpha_t}\|y - x_t\|^2 \right\} \quad (2.7)$$

Crucially the models are assumed to be accurate in the following sense:

$$\mathbb{E}_z[f_x(x,z)] = f(x) \quad \text{and} \quad \mathbb{E}_z[f_x(y,z)] \leq f(y) + \frac{\tau}{2}\|y-x\|^2.$$

Under these two assumptions, the convergence guarantees of the method (2.7) directly parallel that of the basic subgradient method: the expected norm of the Moreau envelope's gradient tends to zero at the rate $O(T^{-1/4})$. The most important algorithms that fit into this framework are the stochastic proximal gradient, clipped gradient,[2] proximal point, and proximal-linear methods. Closely related works include [27], which studies asymptotic convergence guarantees, and [4, 5], which provide intriguing theoretical justifications for using tighter models than linear when designing algorithms.

[2] introduced in [5]

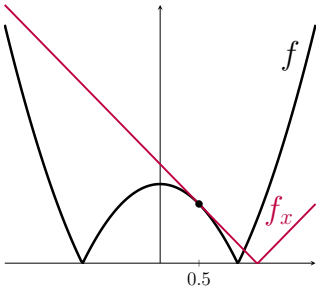


Figure 4: One-sided model: $f(x) = |x^2 - 1|$, $f_{0.5}(y) = |1.25 - y|$

Proximally smooth constraint sets. Weakly convex functions have convex domains, and hence, cannot model nonconvex constraint sets. This is an important issue: one is often interested in optimizing nonsmooth functions—even those that are convex—over nonconvex sets, for example, over embedded submanifolds or over sets cut out by nonconvex functional constraints [2, 57]. The work [23] generalized Algorithm (2.7) to functions φ whose domains are *proximally smooth sets* [18]—a broad class that includes closed convex sets, sublevel sets of weakly convex functions, and compact C^2-submanifolds of $\mathbb{R}^d$.[3]

The Algorithms developed in [23] draw on core techniques of manifold optimization [3] and nonlinear programming [44]. Namely, letting $\mathcal{X}$ denote a proximally smooth constraint set, incorporated in (2.6) as $r(x) = \delta_{\mathcal{X}}$, the method replaces (2.7) with two simpler steps: the first optimizes the model function over a local approximation of $\mathcal{X}$, while the second "retracts" this iterate back to $\mathcal{X}$. Somewhat surprisingly, under natural conditions, the algorithm continues to drive the expectation of the Moreau envelope's gradient to zero at the rate of $O(T^{-1/4})$. The concurrent and complementary work [37] provides a systematic treatment of (stochastic) Riemannian subgradient methods, while the work [40] treats weakly convex minimization with weakly convex functional constraints. Another line of work [38,39,50] extends the Moreau envelope technique to minimax problems

$$\min_x \Phi(x) := \max_{y\in\mathcal{Y}} \varphi(x,y)$$

where $\varphi(x,y)$ is weakly convex in x and concave in y and $\mathcal{Y}$ is a convex set. The function Φ is then weakly convex as well, and therefore the Moreau envelope gradient $\|\nabla\Phi_\lambda(x)\|$ remains a meaningful measure of stationarity. In this setting, one cannot evaluate the inner maximization problem in closed form, and instead (sub)gradient "descent-ascent" algorithms have been developed. Surprisingly, the measure $\|\nabla\Phi_\lambda(x_t)\|$ still tends to zero along the iterate sequence, at the rate of $O(T^{-1/6})$.

[3] Proximally smooth sets have appeared under a variety of names in the literature, including *sets with positive reach* [31] and *uniformly prox-regular sets* [48]. Proximal smoothness was systematically studied in [18] with the view towards optimization theory, though the core definition dates back to Federer [31].

3 Subgradient methods beyond weak convexity

As the previous sections have illustrated, the weakly convex problem class is broadly applicable and yet allows for the design and analysis of efficient algorithms. There are however many simple functions that are not weakly convex, such as $\varphi(x,y) = (|x| - |y|)^2$ and $\varphi(x) = (1 - \max\{x, 0\})^2$. It is not only toy examples, however, that are not weakly convex, but the entire class of deep neural networks with nonsmooth activation functions (e.g., ReLU). Let us look at two concrete examples of non-weakly convex functions that play an important role in applications.

Example 3.1 (Max-affine regression)**.** Consider the regression model

$$y = \max_{1\le j\le k} (\langle a_j, x\rangle + b_j) + \epsilon,$$

where y is a univariate response, $x \in \mathbb{R}^d$ is a vector of covariates, and ϵ is a noise vector. The problem of max-affine regression is to estimate the parameters $\{(a_i, b_i)\}_{i=1}^k$ from independent observations $(x_1, y_1), \ldots, (x_n, y_n)$. Finding the maximum likelihood estimator amounts to solving the problem

$$\min_{(a_j,b_j)} \ \frac{1}{n}\sum_{j=1}^{n} h\left(\max_{1\le j\le k} (\langle a_j, x\rangle + b_j) - y_j\right),$$

where the loss h is determined by the distribution of the noise ϵ. It is straightforward to see that this problem is not weakly convex in general.

Example 3.2 (Deep networks)**.** Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a set of data points in $\mathbb{R}^d \times \mathbb{R}$. A deep neural network loss $\varphi(w; x_j, y_j)$ corresponding to a data point (x_j, y_j) is defined recursively as:

$$\begin{aligned} &a_0 = x_j,\\ &a_i = \rho_i(V_i(w)a_{i-1}) \ \ \forall i = 1, \ldots, L,\\ &\varphi(w; x_j, y_j) = \ell(y_j, a_L), \end{aligned}$$

where $V_i(\cdot)$ are linear maps into the space of matrices, $\ell(\cdot;\cdot)$ is any loss function, and ρ_i are any activation functions applied coordinate wise. Typical examples of losses are the logistic $\ell(y; z) = \log(1 + e^{-yz})$, hinge $\ell(y; z) = \max\{0, 1 - yz\}$, absolute deviation $\ell(y; z) = |y - z|$, and the square loss $\ell(y; z) = \frac{1}{2}(y - z)^2$. Typical activation functions are $\log t$, $\exp(t)$, $\max(0, t)$, or $\log(1 + e^t)$. The task of training a deep network then amounts to the optimization problem

$$\min_{w} \ \frac{1}{n}\sum_{j=1}^{n} \varphi(w; x_j, y_j). \tag{3.1}$$

Whenever the activation functions ρ_i are nonsmooth—the typical setting in practice—the optimization problem (3.1) is not weakly convex.

The approach we follow in this section is based on a continuous time perspective explored in the paper [21], and is orthogonal to the seminal work of Norkin [45].[4] The paper [45] introduced a class of functions called *generalized differentiable*, which subsumes Examples 3.1 and 3.2, and proved convergence of the subgradient method for such functions under an additional Sard-type assumption. The latter assumption holds automatically for Whitney-stratifiable functions, which we discuss in this section.

3.1 A path through continuous time

As we have seen, the Moreau envelope of a weakly convex problem serves as an approximate Lyapunov function for the subgradient method. This is no longer the case outside of the weakly convex setting. An alternative and appealing approach to understanding the asymptotic behavior of the subgradient method is to pass to continuous time where a Lyapunov function may be more apparent. For the sake of simplicity, we only focus on the deterministic subgradient method; all results mentioned here (Section 3) extend to the stochastic setting.

To formally describe the passage to continuous time, let us abstract away from optimization and instead consider the task of solving the inclusion

$$0 \in G(x). \tag{3.2}$$

for a set-valued map $G\colon \mathbb{R}^d \rightrightarrows \mathbb{R}^d$. Assume throughout that G is locally bounded, convex-valued, and has a closed graph. The reader should keep in mind the most important example $G = -\partial\varphi$, where φ is a locally Lipschitz continuous function. In this case, the solutions of (3.2) are precisely the critical points of φ.

The main strategy now for studying the long-term behavior of the discrete process

$$x_{k+1} = x_k + \alpha_k y_k \qquad \text{with} \ \ y_k \in G(x_k). \tag{3.3}$$

is to link its behavior with absolutely continuous solutions $z\colon \mathbb{R}_+ \to \mathbb{R}^d$ of the differential inclusion

$$\dot{z}(t) \in G(z(t)) \qquad \text{for a.e. } t \ge 0. \tag{3.4}$$

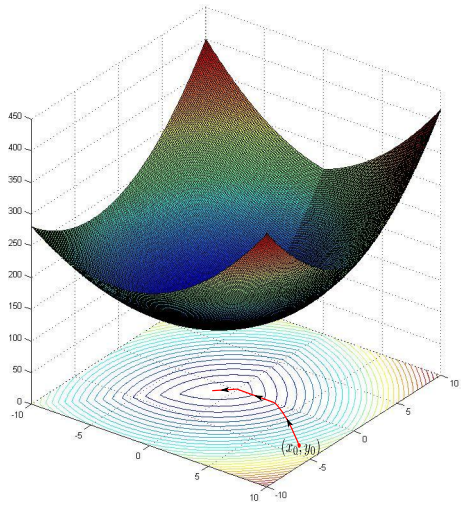
To formalize this viewpoint, we follow the seminal work of [7] and the monograph [12]. Define the time points $t_0 = 0$ and $t_m = \sum_{k=1}^{m-1} \alpha_k$, for $m \ge 1$. Let $x(\cdot)$ now be the linear interpolation of the discrete path:

$$x(t) := x_k + \frac{t - t_k}{t_{k+1} - t_k}(x_{k+1} - x_k) \qquad \text{for } t \in [t_k, t_{k+1}). \tag{3.5}$$

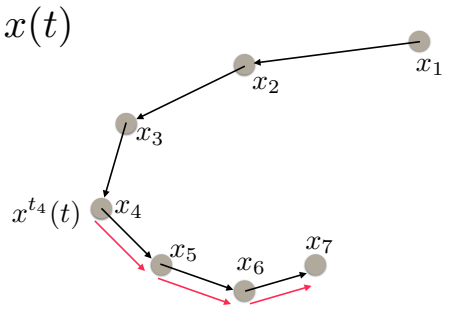
Since we are interested in the asymptotic behavior of trajectories, for each time $\tau \ge 0$ define the time-shifted curve $x^\tau(\cdot) = x(\tau + \cdot)$; see Figure 5b for an illustration.

The following theorem shows that under mild assumptions, the curves $x^\tau(\cdot)$ approximate trajectories of the differential inclusion (3.4) within the space of continuous curves $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$, equipped with the topology of uniform convergence on compact sets.

[4] see also the review article [30].



(a) Subrgadient curve $\dot{z}(t) \in -\partial\varphi(z(t))$.



(b) Time-shifted curves

Figure 5: Dynamics in continuous time

Theorem 3.1 (Functional approximation [7, 12]). *Suppose that the iterates x_t are bounded and that the sequence $\alpha_k > 0$ satisfies $\sum_{i=1}^{k} \alpha_k = \infty$. Then for any sequence $\{\tau_k\}_{k=1}^{\infty} \subseteq \mathbb{R}_+$, the set of functions $\{x^{\tau_k}(\cdot)\}$ is relatively compact in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$. If in addition $\tau_k \to \infty$ as $k \to \infty$, all limit points $z(\cdot)$ of $\{x^{\tau_k}(\cdot)\}$ in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ are trajectories of the differential inclusion (3.4).*

Henceforth, in light of the theorem, we always assume that the sequence α_k satisfies $\sum_{k=1}^{\infty} \alpha_k = \infty$.

Recall that the ultimate goal is to find conditions guaranteeing that every limit point $\bar{x}$ of the sequence $\{x_k\}$, produced by the recursion (3.3), satisfies the desired inclusion (3.2). Making such a leap rigorous typically relies on combining the asymptotic convergence guarantee of Theorem 3.1 with existence of a Lyapunov-like function $\mathcal{L}(\cdot)$ for the continuous dynamics. Let us therefore introduce the following assumption.

Assumption A (Lyapunov condition). *There exists a continuous function $\mathcal{L}\colon \mathbb{R}^d \to \mathbb{R}$, which is bounded from below, and satisfies the following two properties.*

1. **(Weak Sard)** *For a dense set of values $r \in \mathbb{R}$, the intersection $\mathcal{L}^{-1}(r) \cap G^{-1}(0)$ is empty.*
2. **(Descent)** *Whenever $z\colon \mathbb{R}_+ \to \mathbb{R}^d$ is a trajectory of the differential inclusion (3.4) and $0 \notin G(z(0))$, there exists a real $T > 0$ satisfying*

$$\mathcal{L}(z(T)) < \sup_{t\in[0,T]} \mathcal{L}(z(t)) \leq \varphi(z(0)).$$

The weak Sard property is reminiscent of the celebrated Sard's theorem in real analysis. Indeed, consider the classical setting $G = -\nabla\varphi$ for a smooth function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$. Then the weak Sard property stipulates that the set of noncritical values of φ is dense in $\mathbb{R}$. By Sard's theorem, this is indeed the case, as long as φ is C^d smooth. Indeed, Sard's theorem guarantees the much stronger property that the set of noncritical values has full measure. The descent property says that $\mathcal{L}$ eventually strictly decreases along the trajectories of the differential inclusion $\dot{z}(t) \in G(z(t))$ emanating from any non-equilibrium point.

Theorem 3.2 ([7]). *Suppose that Assumption A holds. Then every limit point of $\{x_k\}_{k\geq 1}$ lies in $G^{-1}(0)$ and the function values $\{\mathcal{L}(x_k)\}_{k\geq 1}$ converge.*

3.2 The chain rule along paths

In summary, the asymptotic behavior of the subgradient method can be understood by passing to continuous time if one can verify Assumption A for the set-valued map $G = -\partial\varphi$. With this in mind, a natural candidate for the Lyapunov function is φ itself. Let us put aside for the moment the weak Sard property and focus on the descent property. To this end, we introduce the following intuitive sufficient condition.

Definition 3.3 (Chain rule). Consider a locally Lipschitz function φ on $\mathbb{R}^d$. We will say that φ *admits a chain rule* if for any absolutely continuous curve $z\colon \mathbb{R}_+ \to \mathbb{R}^d$, equality

$$(\varphi \circ z)'(t) = \langle \partial\varphi(z(t)), \dot{z}(t)\rangle \qquad \text{holds for a.e. } t \geq 0.$$

Functions that admit a chain rule automatically satisfy the descent property.

Theorem 3.4 ([21, 25]). *Consider a locally Lipschitz function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ that admits a chain rule. Then $\mathcal{L} = \varphi$ satisfies the descent property (Assumption A(2)) with $G = -\partial\varphi$.*

Convex functions, and more generally those that are "subdifferentially regular", are well-known to admit the chain rule. The two problems outlined in Examples 3.1 and 3.2 however do not fall into these classes. Nonetheless, they do satisfy the chain rule but for a very different reason, which we explain next.

3.3 Stratifiable functions

In this section, we focus on a broad class of functions—virtually exhaustive in applications—that satisfy a chain rule. Roughly speaking, we will call a function *Whitney C^p-stratifiable* if its graph can be decomposed into C^p-smooth

manifolds that fit together in a regular pattern. As an illustration, Figure 6 exhibits such a decomposition of a set in $\mathbb{R}^3$. A formal definition of Whitney C^p-stratifiable can be found in [9, 21], for example.

Semi-algebraic functions comprise the most important subclass of Whitney C^p-stratifiable functions for any finite p. More generally, functions definable in an o-minimal structure—a far-reaching axiomatization of semi-algebraicity [56]—are Whitney C^p-stratifiable. Most importantly, nonsmooth deep neural networks (Example 3.2) built from definable losses $\ell(\cdot, a)$ and definable activation functions ρ_i—such as ReLU, quadratics t^2, hinge losses $\max\{0, t\}$, and SoftPlus $\log(1 + e^t)$ functions—are themselves definable. Therefore convergence guarantees for the subgradient method on Whitney stratifiable functions directly translate to definable deep networks.

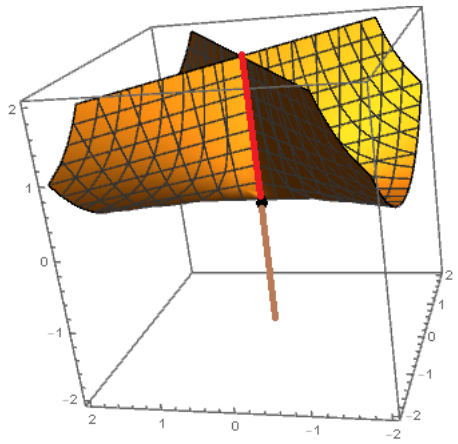


Figure 6: Whitney umbrella ($x^2 = zy^2$) decomposed into strata: origin, positive z-axis, negative z-axis, and their complement.

The following theorem shows that Whitney stratifiable functions automatically admit the chain rule and moreover satisfy the weak Sard property of Assumption A. The Sard result was proved in [9], while the chain rule was established in [21, 25] using the projection formula of [9]. It is worthwhile to mention that Sard type result holds more generally for any set-valued map with a stratifiable graph; see the original work [35] or the monograph [36, §8.4].

Theorem 3.5 (Chain rule and Sard)**.** *Any locally Lipschitz function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ that is Whitney C^p-stratifiable admits a chain rule. Moreover, the set of critical values of φ has zero measure, as long as $p \geq d$. Therefore, in this case, Assumption A holds for $G = -\partial\varphi$.*

Putting Theorems 3.2, 3.4, and 3.5 together immediately yields the following.

Corollary 3.6. *Let $\varphi\colon \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz function that is C^d-stratifiable. Suppose that the iterates $\{x_k\}_{k\geq 1}$ produced by (3.3) are bounded and suppose that Assumption A holds with $G = -\partial\varphi$. Then every limit point of the iterates $\{x_k\}_{k\geq 1}$ is critical for φ and the function values $\{\varphi(x_k)\}_{k\geq 1}$ converge.*

3.4 Further results

Chain rule, conservative maps, and automatic differentiation. The continuous time perspective heavily relied on the validity of the chain rule (Definition 3.3) for the subdifferential of stratifiable functions. In essense, this was the only important property of the Clarke subdifferential used. Indeed, the authors of [10] show that the chain rule can be used to *define* generalized differentiation. Namely, call a set-valued map D_φ a *conservative subdifferential* of φ if it satisfies the mean value theorem in integral form

$$\varphi(\gamma(1)) - \varphi(\gamma(0)) = \int_0^1 \langle D_\varphi(\gamma(t)), \dot\gamma(t)\rangle\, dt$$

for any absolutely continuous curve $\gamma\colon [0, 1] \to \mathbb{R}^d$. Conservative subdifferentials are systematically studied in [10]. Importantly, conservative subdifferentials satisfy an *exact* calculus, which is in sharp contrast to the Clarke subdifferential. In particular, it is immediate from the definition that $D_f + D_g$ is a conservative subdifferential of $f + g$.

The most important example of conservative subdifferentials arises from the Automatic differentiation technique commonly used in deep learning. To motivate the discussion, recall that implementation of the subgradient method requires a mechanism to efficiently compute a Clarke subgradient $y_t \in \partial\varphi(x_t)$. There are instances when this is not entirely justified. For example, when training deep neural networks, computing a subgradient amounts to subdifferentiating a long composition of functions $\varphi = f_1 \circ f_2 \circ \ldots \circ f_L$. The way this is done in practice is by appealing to automatic differentiation techniques, which implicitly compute an update direction from the subdifferentials ∂f_i by "formally" applying a chain rule, as if the function f_i were differentiable. Herein lies a conceptual disconnect between theory and practice because the subdifferential of nonsmooth functions in general does not satisfy formal chain or sum rules with equality. Nonetheless, the authors of [10, 11] prove that the update direction defined in this way furnishes a conservative subdifferential and therefore the following is true.

Theorem 3.7 (Informal)**.** *Consider a function $\varphi = f_1 \circ f_2 \circ \ldots \circ f_L$ where each $f_i\colon \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz continuous and admits a chain rule. Let $G\colon \mathbb{R}^d \rightrightarrows \mathbb{R}$ be a set-valued map defined by formally applying the chain rule of subdifferentials to the decomposition of φ. Fix a sequence $\alpha_i > 0$ satisfying $\sum_{i=1}^\infty \alpha_i = \infty$ and consider the iterates:*

$$x_{t+1} = x_t - \alpha_t y_t \qquad \text{where } y_t \in G(x_t).$$

Then as long as the sequence $\{x_t\}$ is bounded, each of its limit points x satisfies the inclusion $0 \in G(x)$.

An unsatisfactory aspect of Theorem 3.7 is that in principle a subgradient method implemented using automatic differentiation can generate points that are not Clarke critical. Seeking to avoid such pathologies, the authors of [10,11] show remarkably that for a large class of deep neural networks, as long as the initialization and the step-sizes are sufficiently random, every limit point of the method will be Clarke critical. A parallel approach appears in [8].

Theorem 3.8 (Informal). *Consider applying the stochastic gradient method implemented through automatic differentiation, with stepsizes chosen as $\alpha_t = c\gamma_t$ where $c \in (0,1)$ and $\gamma_t = \mathcal{O}(1/\log(t))$. Then for a virtually exhaustive class of activation functions (see [11] for a formal definition), for almost every $c \in (0,1)$ and almost all $x_0 \in \mathbb{R}^d$, the stochastic subgradient method produces limit point x that satisfy the inclusion $0 \in \partial\varphi(x)$ almost surely.*

Finite time guarantees. The passage to continuous time provides an attractive framework for understanding the asymptotic behavior of subgradient methods. The disadvantage of this approach is that it does not yield insight into finite time guarantees. The recent work [58] establishes finite time guarantees for a modified subgradient method, which is closely related to the descent method of Goldstein [34] and the gradient sampling algorithm of Burke, Lewis, and Overton [14]. The main construction that used is the Goldstein subdifferential introduced in [34].

Definition 3.9 (Goldstein subdifferential). Consider a locally Lipschitz function $\varphi\colon \mathbb{R}^d \to \mathbb{R}$, a point $x \in \mathbb{R}^d$, and a parameter $\delta > 0$. The *Goldstein subdifferential* of φ at x is the set

$$\partial_\delta \varphi(x) = \operatorname{conv}\Big(\bigcup_{x \in B_\delta(x)} \partial\varphi(x) \Big).$$

Thus the Goldstein subdifferential of φ at x is simply the convex hull of the union of all Clarke subgradients at points in a δ-ball around x. Famously, Goldstein [34] showed that one can significantly decrease the value of φ by taking a step in the direction of the minimal norm element of $\partial_\delta\varphi(x)$. Throughout the rest of the section, we fix $\delta \in (0,1)$.

Theorem 3.10 (Uniform decrease). *Fix a point x and let g be a minimal norm element of $\partial_\delta\varphi(x)$. Then as long as $g \neq 0$, the estimate holds:*

$$\varphi\left(x - \delta\frac{g}{\|g\|}\right) \leq \varphi(x) - \delta\|g\|.$$

Theorem 3.10 immediately motivates a conceptual descent algorithm, which repeats:

$$x_{t+1} = x_t - \delta\frac{g_t}{\|g_t\|} \qquad \text{where} \qquad g_t = \operatorname*{arg\,min}_{g \in \partial_\delta\varphi(x)} \|g\|. \quad (3.6)$$

Theorem 3.10 trivially guarantees that the *stationarity condition* $\min_{t=1,\ldots,T} \|g_t\| \leq \epsilon$ holds after

$$T = \mathcal{O}\left(\frac{\varphi(x_0) - \min\varphi}{\delta\epsilon}\right) \quad \text{iterations.}$$

Since evaluating the minimal norm element of $\partial_\delta\varphi(x)$ is impossible in general, the descent method cannot be applied directly. Nonetheless it does serve as a guiding principle for implementable algorithms. Notably, the gradient sampling algorithm [14] in each iteration forms polyhedral approximations K_t of $\partial_\delta\varphi(x_t)$ by sampling gradients in the ball $B_\delta(x)$ and computes search directions $g_t \in \arg\min_{g \in K_t} \|g\|$. The number of gradient computations required by gradient sampling algorithms, however, scales linearly with the dimension of the ambient space; see [13].

The recent paper [58] shows, remarkably, that for any $x \in \mathbb{R}^d$ one can find an approximate minimal norm element of $\partial_\delta\varphi(x)$ using a number of subgradients that is independent of the dimension.[5] As a consequence the following is true.

Theorem 3.11 (Informal). *Let φ be a Lipschitz continuous and directionally differentiable function and set $\Delta = \varphi(x_0) - \min\varphi$. There exists an algorithm that with probability $1-\gamma$ will find a point x satisfying $\operatorname{dist}(0, \partial_\delta\varphi(x)) \leq \epsilon$ using at most $\mathcal{O}\left(\frac{\Delta L^2}{\delta\epsilon^3}\log\left(\frac{4\Delta}{\gamma\delta\epsilon}\right)\right)$ subgradient evaluations.*

It is natural to ask whether for general Lipschitz functions one may efficiently find some point x for which there exists $y \in \mathbb{B}_\delta(x)$ satisfying $\operatorname{dist}(0, \partial\varphi(y)) \leq \epsilon$. This is a much stronger requirement than $\operatorname{dist}(0, \partial_\delta\varphi(x)) \leq \epsilon$, and was exactly the guarantee of subgradient methods on weakly convex functions in Corollary 2.1. The paper [52] shows that for general Lipschitz continuous functions, the number of subgradient computations required to achieve this goal by any algorithm must scale with the dimension of the ambient space.

4 Conclusion

The subgradient method has long been a useful algorithm for minimizing nonsmooth functions, and is gaining renewed prominence due to large-scale applications in machine learning. It is therefore crucial to determine whether and when the subgradient method possesses theoretical guarantees. In this work we outlined two broad problem classes, both amenable to theoretical analysis: weakly convex and tame problems. For weakly convex functions, the subgradient method satisfies strong complexity guarantees: the gradient of the Moreau envelope of the loss function tends to zero at a controlled rate. For tame problems, the situation is much more subtle: while asymptotic guarantees hold, available complexity guarantees are much weaker. An intriguing open question is whether such complexity guarantees can be strengthened.

References

[1] E. Abbe, A.S. Bandeira, A. Bracher, and A. Singer. Decoding binary node labels from censored edge measurements: phase transition and efficient recovery. *IEEE Trans. Network Sci. Eng.*, 1(1):10–22, 2014.

[2] P.-A. Absil and S. Hosseini. A collection of nonsmooth Riemannian optimization problems. In *Nonsmooth Optimization and Its Applications*, pages 1–15. Springer, 2019.

[3] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds.* Princeton University Press, 2009.

[4] H. Asi and J.C. Duchi. The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019.

[5] Strictly speaking, the procedure [58] requires to compute a "particular" subgradient v at each query point x satisfying $f'(x,v) = \langle v, x\rangle$.

[5] H. Asi and J.C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

[6] A.S. Bandeira, N. Boumal, and V. Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, June 23-26, 2016*, pages 361–382, 2016.

[7] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

[8] Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *arXiv preprint arXiv:2005.08513*, 2020.

[9] J. Bolte, A. Daniilidis, A.S. Lewis, and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.

[10] J. Bolte and E. Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, pages 1–33, 2020.

[11] J. Bolte and E. Pauwels. A mathematical model for automatic differentiation in machine learning. *arXiv preprint arXiv:2006.02080*, 2020.

[12] V.S. Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[13] J.V. Burke, F.E. Curtis, A.S. Lewis, M.L. Overton, and L.E.A. Simões. Gradient sampling methods for nonsmooth optimization. In *Numerical Nonsmooth Optimization*, pages 201–225. Springer, 2020.

[14] J.V. Burke, A.S. Lewis, and M.L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

[15] E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *J. ACM*, 58(3):Art. 11, 37, 2011.

[16] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A.S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, 21(2):572–596, 2011.

[17] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. *arXiv preprint arXiv:1904.10020*, 2019.

[18] F.H. Clarke, R.J. Stern, and P.R. Wolenski. Proximal smoothness and the lower-C^2 property. *J. Convex Anal.*, 2(1-2):117–144, 1995.

[19] A. Daniilidis and D. Drusvyatskiy. Pathological subgradient dynamics. *SIAM Journal on Optimization*, 30(2):1327–1338, 2020.

[20] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[21] D. Davis, D. Drusvyatskiy, S. Kakade, and J.D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

[22] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *IMA Journal of Numerical Analysis*, 40(4):2652–2695, 2020.

[23] D. Davis, D. Drusvyatskiy, and Z. Shi. Stochastic optimization over proximally smooth sets. *arXiv preprint arXiv:2002.06309*, 2020.

[24] D. Drusvyatskiy. The proximal point method revisited. *SIAG/OPT Views and News*, 26(1), 2018.

[25] D. Drusvyatskiy, A.D. Ioffe, and A.S. Lewis. Curves of descent. *SIAM Journal on Control and Optimization*, 53(1):114–138, 2015.

[26] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1-2):503–558, 2019.

[27] J.C. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

[28] J.C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, 8(3):471–529, 2019.

[29] Y.C. Eldar and S. Mendelson. Phase retrieval: stability and recovery guarantees. *Appl. Comput. Harmon. Anal.*, 36(3):473–494, 2014.

[30] Yu.M. Ermol'ev and V.I. Norkin. Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization. *Cybernetics and Systems Analysis*, 34(2):196–215, 1998.

[31] H. Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

[32] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[33] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[34] A.A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

[35] A.D. Ioffe. Critical values of set-valued maps with stratifiable graphs. Extensions of Sard and Smale-Sard theorems. *Proceedings of the American Mathematical Society*, 136(9):3111–3119, 2008.

[36] A.D. Ioffe. Variational analysis of regular mappings. *Springer Monographs in Mathematics. Springer, Cham*, 2017.

[37] X. Li, S. Chen, Z. Deng, Q. Qu, Z. Zhu, and A.M.C. So. Nonsmooth optimization over stiefel manifold: Riemannian subgradient methods. *arXiv preprint arXiv:1911.05047*, 2019.

[38] T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6083–6093, Virtual, 13–18 Jul 2020. PMLR.

[39] T. Lin, C. Jin, and M.I. Jordan. Near-optimal algorithms for minimax optimization. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2738–2779. PMLR, 09–12 Jul 2020.

[40] R. Ma, Q. Lin, and T. Yang. Proximally constrained methods for weakly convex optimization with weakly convex constraints. *arXiv preprint arXiv:1908.01871*, 2019.

[41] J.-J. Moreau. Proximité and dualité in a Hilbert space. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[42] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[43] A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[44] J. Nocedal and S.J. Wright. *Numerical optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[45] V.I. Norkin. Stochastic generalized-differentiable functions in the problem of nonconvex nonsmooth stochastic optimization. *Cybernetics*, 22(6):804–809, 1986.

[46] E.A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

[47] E.A. Nurminskii. Minimization of nondifferentiable functions in the presence of noise. *Cybernetics*, 10(4):619–621, Jul 1974.

[48] R. Poliquin and R.T. Rockafellar. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348(5):1805–1838, 1996.

[49] R.A. Poliquin and R.T. Rockafellar. Amenable functions in optimization. In *Nonsmooth optimization: methods and applications (Erice, 1991)*, pages 338–353. Gordon and Breach, Montreux, 1992.

[50] H. Rafique, M. Liu, Q. Lin, and T. Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

[51] R.T. Rockafellar. Favorable classes of Lipschitz-continuous functions in subgradient optimization. In *Progress in nondifferentiable optimization*, volume 8 of *IIASA Collaborative Proc. Ser. CP-82*, pages 125–143. Int. Inst. Appl. Sys. Anal., Laxenburg, 1982.

[52] O. Shamir. Can we find near-approximately-stationary points of nonsmooth nonconvex functions? *arXiv preprint arXiv:2002.11962*, 2020.

[53] M. Shiota. *Geometry of subanalytic and semialgebraic sets*, volume 150. Springer Science & Business Media, 2012.

[54] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.*, 30(1):20–36, 2011.

[55] L. Van den Dries. *Tame topology and o-minimal structures*, volume 248. Cambridge University Press, 1998.

[56] L. Van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Math. J*, 84(2):497–540, 1996.

[57] T. Yang. Advancing non-convex and constrained learning: Challenges and opportunities. *AI Matters*, 5(3):29–39, December 2019.

[58] J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11173–11182, Virtual, 13–18 Jul 2020. PMLR.

Bulletin

Email items to siagoptnews@lists.mcs.anl.gov *for consideration in the bulletin of forthcoming issues.*

1 Event Announcements

1.1 SIAM Conference on Optimization

Although originally scheduled for May 26-29 in Hong Kong, because of the COVID-19 pandemic SIOPT has been moved to 2021. It will be held in July 20-23 in Spokane, Washington, USA, jointly with four other meetings: the SIAM Annual Meeting (AN21); the SIAM Conference on Applied and Computational Discrete Algorithms (ACDA21); the SIAM Conference on Control and Its Applications (CT21); and the SIAM Conference on Discrete Mathematics (DM21). More information at www.siam.org/conferences/cm/conference/op21.

1.2 IPCO

The 22nd Conference on Integer Programming and Combinatorial Optimization (IPCO XXII) will take place on 19–21 May, 2021, at the Georgia Institute of Technology, Atlanta, Georgia, USA. It will be preceded by a Summer School, to be held on 17 and 18 May.

For details visit https://sites.gatech.edu/ipco-2021.

1.3 IFORS

The 22nd Conference of the International Federation of Operational Research Societies (IFORS 2021) will take place 22-27 August 2021 at Hanyang University, Seoul, South Korea. It was originally scheduled for 2020 but was also moved to 2021 due to the pandemic.

See http://www.ifors2020.kr for more information.

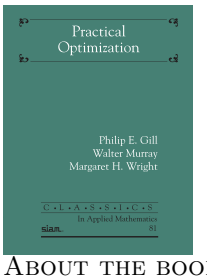
1.4 MOPTA

The MOPTA 2021 conference will be held at Lehigh University, 2-4 August 2021. MOPTA aims at bringing together a diverse group of people from both discrete and continuous optimization, working on both theoretical and applied aspects. There will be a small number of invited talks from distinguished speakers and contributed talks, spread over three days.

For details visit http://coral.ie.lehigh.edu/~mopta

2 Books

2.1 Practical Optimization



By Philip E. Gill, Walter Murray, and Margaret H. Wright

Publisher: SIAM

ISBN: 978-1-611975-59-8

Published: 2019

https://epubs.siam.org/doi/book/10.1137/1.9781611975604

About the book:

In the intervening years since this book was published in 1981, the field of optimization has been exceptionally lively. This fertility has involved not only progress in theory, but also faster numerical algorithms and extensions into unexpected or previously unknown areas such as semidefinite programming. Despite these changes, many of the important principles and much of the intuition can be found in this Classics version of Practical Optimization.

This book

- provides model algorithms and pseudocode, useful tools for users who prefer to write their own code as well as for those who want to understand externally provided code;
- presents algorithms in a step-by-step format, revealing the overall structure of the underlying procedures and thereby allowing a high-level perspective on the fundamental differences; and
- contains a wealth of techniques and strategies that are well suited for optimization in the twenty-first century and particularly in the now-flourishing fields of data science, "big data," and machine learning.

Audience: Practical Optimization is appropriate for advanced undergraduates, graduate students, and researchers interested in methods for solving optimization problems.

3 Other Announcements

3.1 Martin Grötschel

Martin Grötschel has been awarded the Cantor Medal, the highest scientific award of the German Mathematicians' Association (DMV), for 2021. Congratulations Martin!

3.2 Arkadi Nemirovski

Arkadi Nemirovski has been elected as member of the United States' National Academy of Sciences for his contributions to continuous optimization. Congratulations Arkadi!

Chair's Column

Katya Scheinberg, SIAG/OPT Chair
Cornell University, Ithaca, NY 18015-1582, USA
katyas@cornell.edu
https:/www.orie.cornell.edu/faculty-directory/katya-scheinberg

It is my privilege to write this column for the first time as the new Chair of SIAM Activity Group on Optimization. The new team stepped in on January 1st of 2020 and in addition to myself includes Sam Burer as the Vice Chair, Jeff Linderoth as the Program Director and Stefan Wild as the Secretary. We feel very honored to have been elected to lead our SIAG but also downhearted because of the effect which Covid-19 has had on our community and in particular on the postponement and relocation of our flagship conference (OP 20) which was to be held in Hong Kong last Spring. We hoped to see many of you there during talks, social events and SIAG business meeting.

Tamás Terlaky and Defeng Sun who are organizing committee co-Chairs of the OP 20 conference have been hard at work since the end of 2019 trying to find the best solution for the SIAM on Optimization conference. Defeng and the organizing committee have put in a huge effort towards ensuring a successful and safe conference in Hong Kong, however, the new reality set on the whole world and plans had to be changed. We very much hope to rely on the hospitality and organizational skills of the Hong Kong team again in the near future.

After many deliberations it has been decided that the next instantiation of the OP 20 conference is to be moved and to be collocated with the SIAM Annual meeting. Now named OP 21, the conference is scheduled to be held in Spokane, Washington July 20-23rd 2021. Many of you have seen the call for minisymposium proposals and abstract submissions. We hope to have a successful and vibrant conference in whatever form will be feasible in July. The conference website includes information about the possible modes of the conference and related logistical details.

I would now like to take the opportunity to thank Tamás Terlaky (Chair) and the previous team of officers, Andreas Wächter (Vice Chair), Michael Friedlander (Program Director) and Jim Luedtke (Secretary) for their effective and tireless leadership over the past three years. Under their leadership the SIAG has flourished considerably gaining many new members. A new Early Career Prize has been established and awarded for the first time this year. While the SIAM Conference on Optimization has been postponed, the SIAG prizes have been awarded without delay. The winners of the SIAG/Optimization Best Paper Prize are Hamza Fawzi (U. of Cambridge), James Saunderson (Monash University), and Pablo Parrilo (MIT) for their paper "Semidefinite Approximations of the Matrix Logarithm." And the winner of the new Early Career Prize is John Duchi (Stanford University). Congratulations to the winners!

I wish you and your loved ones to stay healthy and productive during these difficult times and I hope we will all be seeing each other in person soon. Happy New Year.

Comments from the Editors

This edition of *SIAG/OPT Views and News* came a bit late in a year that was eventful for all of us. We have a great contribution by Damek Davis and Dmitriy Drusvyatskiy, showing us the strength of the subgradient method for two classes of optimization problems: weakly convex problems and tame problems. We hope you enjoy the reading and remind you that all volumes of *Views and News* are available at the online archive: http://wiki.siam.org/siag-op/index.php/View_and_News.

As always, the editors welcome your feedback at siagoptnews@lists.mcs.anl.gov. Suggestions for new issues, comments, and papers are always welcome.

Pietro Belotti, Editor
DEIB, Politecnico di Milano, and FICO Xpress team, pietro.belotti@polimi.it

Somayeh Moazeni, Editor
School of Business, Stevens Institute of Technology, smoazeni@stevens.edu, http://web.stevens.edu/facultyprofile/?id=2041