Stochastic subgradient method converges on tame functions[1]

Damek Davis[2]

School of Operations Research and Information Engineering
Cornell University

[1] Joint work with Dima Drusvyatskiy (UW), Sham Kakade (UW), and Jason Lee (Princeton)
Foundations of Computational Mathematics (2019)

[2] https://people.orie.cornell.edu/dsd95/

Guiding Question

Is there a unified way to understand asymptotic convergence in nonsmooth and nonconvex (stochastic) optimization?

Guiding Question

Is there a unified way to understand asymptotic convergence in nonsmooth and nonconvex (stochastic) optimization?

Why care?

- Stochastic (sub)gradient method is standard option in industry backed solvers (Tensorflow, Pytorch).
- Key data science tasks are nonsmooth and nonconvex (ReLU deep networks).

Guiding Question

Is there a unified way to understand asymptotic convergence in nonsmooth and nonconvex (stochastic) optimization?

Why care?

- Stochastic (sub)gradient method is standard option in industry backed solvers (Tensorflow, Pytorch).
- Key data science tasks are nonsmooth and nonconvex (ReLU deep networks).

Many have contributed.

- Belenkiy, Bertsekas, Burke, Demyanov, Duchi, Ermoliev, Gaivoronski, Goffin, Gupal, Juditsky, Kiwiel, Lan, Lemaréchal, Lewis, Mifflin, Mikhalevich, Nemirovski, Nesterov, Norkin, Nurminskii, Overton, Polyak, Pshenichny, Rubinov, Rucinski, Sagastizábal, Shapiro, Shor, Uryasev....

Problem Class and Algorithms I

Problem. Minimize locally Lipschitz function

$$\min_{x \in \mathbb{R}^d} \; f(x).$$

Problem Class and Algorithms I

Problem. Minimize locally Lipschitz function

$$\min_{x \in \mathbb{R}^d} f(x).$$

Example Algorithm:

- Subgradient method

$$\boxed{\begin{aligned} &\text{Choose } y_k \in \partial f(x_k) \\ &x_{k+1} = x_k - \alpha_k y_k. \end{aligned}}$$

Problem Class and Algorithms I

Problem. Minimize locally Lipschitz function

$$\min_{x \in \mathbb{R}^d} \; f(x).$$

Example Algorithm:

- Subgradient method

$$\boxed{\begin{aligned} &\text{Choose } y_k \in \partial f(x_k) \\ &x_{k+1} = x_k - \alpha_k y_k. \end{aligned}}$$

where ∂f denotes Clarke subdifferential:

$$\partial f(x) = \operatorname{conv}\left\{ \lim_{x_i \to x} \nabla f(x_i) : x_i \to x \text{ in } \operatorname{dom}(\nabla f) \right\}.$$

Problem Class and Algorithms II

Problem.

$$\min_{x \in \mathcal{X}} \ \varphi(x) := f(x) + g(x)$$

Problem Class and Algorithms II

Problem.

$$\min_{x \in \mathcal{X}} \; \varphi(x) := f(x) + g(x)$$

Example Algorithms:

- Proximal subgradient method

$$\boxed{\begin{array}{l} \text{Choose } y_k \in \partial f(x_k) \\ x_{k+1} \in \underset{x \in \mathcal{X}}{\arg\min} \;\; f(x_k) + \langle y_k, x - x_k \rangle + g(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2. \end{array}}$$

Problem Class and Algorithms II

Problem.

$$\min_{x \in \mathcal{X}} \; \varphi(x) := f(x) + g(x)$$

Example Algorithms:

- Proximal subgradient method

$$\boxed{\begin{array}{l} \text{Choose } y_k \in \partial f(x_k) \\ x_{k+1} \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \; f(x_k) + \langle y_k, x - x_k \rangle + g(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2. \end{array}}$$

- Clipped proximal subgradient method if $f \geq 0$ (Duchi-Ruan '18)

$$\boxed{\begin{array}{l} \text{Choose } y_k \in \partial f(x_k) \\ x_{k+1} \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \; [f(x_k) + \langle y_k, x - x_k \rangle]^+ + g(x) + \frac{1}{2\alpha_k} \|x - x_k\|^2. \end{array}}$$

Problem Class and Algorithms III

Problem.

$$\min_{x \in \mathcal{X}} \ \varphi(x) := \mathbb{E}_{z \sim \mathbb{P}} \left[f(x, z) \right] + g(x)$$

Problem Class and Algorithms III

Problem.

$$\min_{x \in \mathcal{X}} \; \varphi(x) := \mathbb{E}_{z \sim \mathbb{P}}\left[f(x, z)\right] + g(x)$$

Example Algorithms:

- Stochastic proximal point

$$\boxed{\begin{array}{l} \text{Sample } z_k \sim \mathbb{P} \\ x_{k+1} \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \; f(x, z_k) + g(x) + \dfrac{1}{2\alpha_k}\|x - x_k\|^2. \end{array}}$$

Problem Class and Algorithms III

Problem.

$$\min_{x \in \mathcal{X}} \; \varphi(x) := \mathbb{E}_{z \sim \mathbb{P}}\left[f(x, z)\right] + g(x)$$

Example Algorithms:

- Stochastic proximal point

$$\boxed{\begin{aligned} &\text{Sample } z_k \sim \mathbb{P} \\ &x_{k+1} \in \operatorname*{arg\,min}_{x \in \mathcal{X}} \; f(x, z_k) + g(x) + \frac{1}{2\alpha_k}\|x - x_k\|^2. \end{aligned}}$$

- And stochastic variants of previous algorithms....

Challenge

Lyapunov analysis:

Challenge

Lyapunov analysis:

- φ convex $\implies$ $\|x_k - x^*\|^2$ almost decreasing (Shor '64)

Challenge

Lyapunov analysis:

- φ convex $\implies$ $\|x_k - x^*\|^2$ almost decreasing (Shor '64)
- φ smooth $\implies$ $\varphi(x_k)$ almost decreasing (Ghadimi-Lan '13)

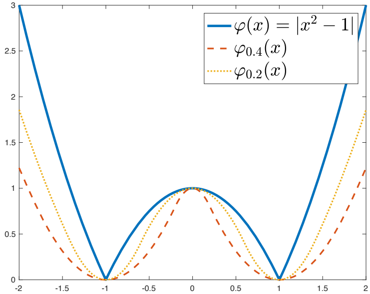
Challenge

Lyapunov analysis:

- φ convex $\implies$ $\|x_k - x^*\|^2$ almost decreasing (Shor '64)
- φ smooth $\implies$ $\varphi(x_k)$ almost decreasing (Ghadimi-Lan '13)
- φ weakly convex $\implies$ Moreau envelope

$$\varphi_\lambda(x) = \inf_y \left\{\varphi(y) + \tfrac{1}{2\lambda}\|y - x\|^2\right\}$$

almost decreasing (D-Drusvyatskiy '18)



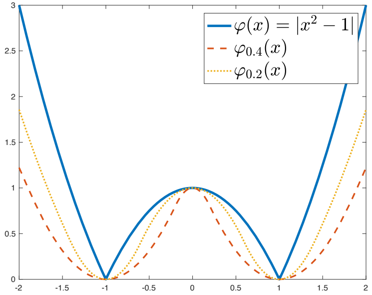
Challenge

Lyapunov analysis:

- φ convex $\implies$ $\|x_k - x^*\|^2$ almost decreasing (Shor '64)
- φ smooth $\implies$ $\varphi(x_k)$ almost decreasing (Ghadimi-Lan '13)
- φ weakly convex $\implies$ Moreau envelope

$$\varphi_\lambda(x) = \inf_y \left\{\varphi(y) + \tfrac{1}{2\lambda}\|y - x\|^2\right\}$$

almost decreasing (D-Drusvyatskiy '18)



Challenge: No clear Lyapunov function in general.

The Differential Inclusion Approach

Define: Composite Gradient

$$G(z) := \partial f(z) + \partial g(z) + N_{\mathcal{X}}(z).$$

The Differential Inclusion Approach

Define: Composite Gradient

$$G(z) := \partial f(z) + \partial g(z) + N_{\mathcal{X}}(z).$$

Two Ingredients in Search for Critical Point: $0 \in G(x)$

The Differential Inclusion Approach

Define: Composite Gradient

$$G(z) := \partial f(z) + \partial g(z) + N_{\mathcal{X}}(z).$$

Two Ingredients in Search for Critical Point: $0 \in G(x)$

- Unifying Principle. Common algorithms are "discretizations" of trajectory $z\colon \mathbb{R}_+ \to \mathbb{R}^d$ of differential inclusion $-\dot{z} \in G(z)$

The Differential Inclusion Approach

Define: Composite Gradient

$$G(z) := \partial f(z) + \partial g(z) + N_{\mathcal{X}}(z).$$

Two Ingredients in Search for Critical Point: $0 \in G(x)$

- Unifying Principle. Common algorithms are "discretizations" of trajectory $z \colon \mathbb{R}_+ \to \mathbb{R}^d$ of differential inclusion $-\dot{z} \in G(z)$
- Lyapunov Assumption.
  - **Strict Descent.** We force φ to be Lyapunov for dynamics.

$$\boxed{\left\{\begin{array}{l} -\dot{z} \in G(z) \text{ a.e.} \\ z(0) \text{ not critical.} \end{array}\right\} \implies \varphi(z(t)) < \varphi(z(0)) \quad \forall t > 0}$$

The Differential Inclusion Approach

Define: Composite Gradient

$$G(z) := \partial f(z) + \partial g(z) + N_{\mathcal{X}}(z).$$

Two Ingredients in Search for Critical Point: $0 \in G(x)$

- Unifying Principle. Common algorithms are "discretizations" of trajectory $z \colon \mathbb{R}_+ \to \mathbb{R}^d$ of differential inclusion $-\dot{z} \in G(z)$
- Lyapunov Assumption.
  - **Strict Descent.** We force φ to be Lyapunov for dynamics.

$$\boxed{\left\{\begin{array}{l} -\dot{z} \in G(z) \text{ a.e.} \\ z(0) \text{ not critical.} \end{array}\right\} \implies \varphi(z(t)) < \varphi(z(0)) \quad \forall t > 0}$$

  - **Weak Sard.** The set of noncritical values is dense in $\mathbb{R}$.

The Differential Inclusion Approach

Define: Composite Gradient

$$G(z) := \partial f(z) + \partial g(z) + N_{\mathcal{X}}(z).$$

Two Ingredients in Search for Critical Point: $0 \in G(x)$

- Unifying Principle. Common algorithms are "discretizations" of trajectory $z\colon \mathbb{R}_+ \to \mathbb{R}^d$ of differential inclusion $-\dot{z} \in G(z)$
- Lyapunov Assumption.
  - **Strict Descent.** We force φ to be Lyapunov for dynamics.

$$\boxed{\left\{\begin{array}{l} -\dot{z} \in G(z) \text{ a.e.} \\ z(0) \text{ not critical.} \end{array}\right\} \implies \varphi(z(t)) < \varphi(z(0)) \quad \forall t > 0}$$

  - **Weak Sard.** The set of noncritical values is dense in $\mathbb{R}$.

Thm: *Lyapunov* $\implies$ *every limit point x^* of $\{x_k\}$ is critical.*

(Kushner-Yin '03, Benaïm-Hofbauer-Sorin '05)

Discretization I

$$\boxed{x_{k+1} = x_k - \alpha_k(y_k + \xi_k).}$$

Assumptions:

Discretization I

$$\boxed{x_{k+1} = x_k - \alpha_k(y_k + \xi_k).}$$

Assumptions:

1. $\sup_{k\geq 0}\{\|x_k\|, \|y_k\|\} < \infty$ a.s.

Discretization I

$$\boxed{x_{k+1} = x_k - \alpha_k(y_k + \xi_k).}$$

Assumptions:

1. $\sup_{k \geq 0}\{\|x_k\|, \|y_k\|\} < \infty$ a.s.
2. Step-size selection

$$\alpha_k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Discretization I

$$\boxed{x_{k+1} = x_k - \alpha_k(y_k + \xi_k).}$$

Assumptions:

1. $\sup_{k \geq 0}\{\|x_k\|, \|y_k\|\} < \infty$ a.s.
2. Step-size selection
$$\alpha_k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$
3. Approximate evaluations
$$x_{k_j} \to x \qquad \Longrightarrow \qquad \frac{1}{n}\sum_{k=1}^{n} y_{k_j} \to G(x).$$

Discretization I

$$\boxed{x_{k+1} = x_k - \alpha_k(y_k + \xi_k).}$$

Assumptions:

1. $\sup_{k \geq 0}\{\|x_k\|, \|y_k\|\} < \infty$ a.s.
2. Step-size selection
$$\alpha_k \geq 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty, \qquad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$
3. Approximate evaluations
$$x_{k_j} \to x \qquad \Longrightarrow \qquad \frac{1}{n}\sum_{k=1}^{n} y_{k_j} \to G(x).$$
4. Noise sequence $\{\xi_k\}$ satisfies
$$\sum_{k=1}^{\infty} \alpha_k \xi_k \text{ exists.}$$

(Kushner-Yin '03, Benaïm-Hofbauer-Sorin '05, Duchi-Ruan '17)

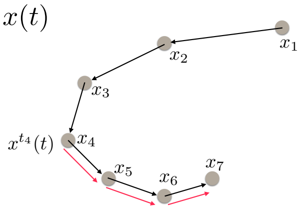
Discretization II

Interpolation.

$x(t)$ linearly interpolates $\{x_k\}$

Time-shifted curve.

$$x^\tau(\cdot) = x(\tau + \cdot).$$



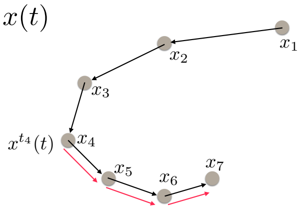
Discretization II

Interpolation.

$x(t)$ linearly interpolates $\{x_k\}$

Time-shifted curve.

$$x^{\tau}(\cdot) = x(\tau + \cdot).$$



Thm: For any $\tau_k \to \infty$, the set $\{x^{\tau_k}(\cdot)\}$ is compact in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$, and all limit points $z(\cdot)$ are arcs satisfying

$$-\dot{z}(t) \in G(z(t)) \qquad \text{for a.e. } t \geq 0.$$

(Kushner-Yin '03, Benaïm-Hofbauer-Sorin '05, Duchi-Ruan '17)

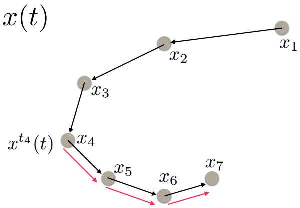
Discretization II

Interpolation.

$x(t)$ linearly interpolates $\{x_k\}$

Time-shifted curve.

$$x^\tau(\cdot) = x(\tau + \cdot).$$



Thm: For any $\tau_k \to \infty$, the set $\{x^{\tau_k}(\cdot)\}$ is compact in $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$, and all limit points $z(\cdot)$ are arcs satisfying

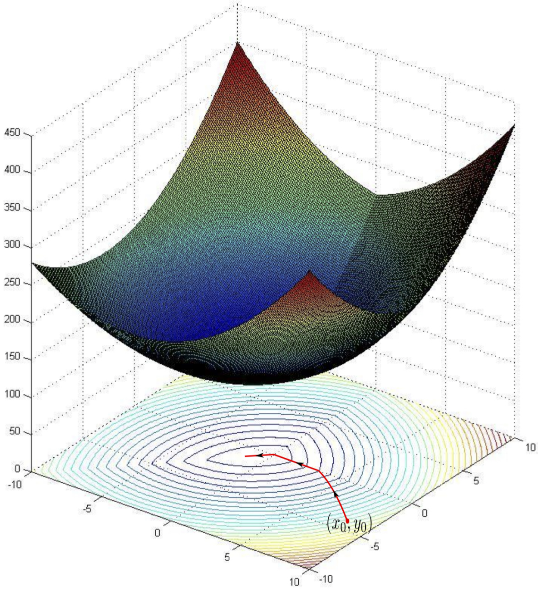
$$-\dot{z}(t) \in G(z(t)) \qquad \text{for a.e. } t \geq 0.$$

(Kushner-Yin '03, Benaïm-Hofbauer-Sorin '05, Duchi-Ruan '17)

Why Matter? If $x_{k_j} \to x^*$, then a limiting arc begins at limit point.

Strict Descent

When does φ strictly decrease along dynamics $-\dot{z} \in G(z)$?



A Sufficient Condition for Strict Descent

Intuition. Dynamics "should" decrease
φ in proportion to square of "gradient."

A Sufficient Condition for Strict Descent

Intuition. Dynamics "should" decrease φ in proportion to square of "gradient."

Sufficient Condition. A chain rule: for any arc z, we have for a.e. t

$$(\varphi \circ z)'(t) = \langle G(z(t)), \dot{z}(t)\rangle$$

A Sufficient Condition for Strict Descent

Intuition. Dynamics "should" decrease φ in proportion to square of "gradient."

Sufficient Condition. A chain rule: for any arc z, we have for a.e. t

$$(\varphi \circ z)'(t) = \langle G(z(t)), \dot{z}(t) \rangle$$

Lemma: Suppose φ admits a chain rule and an arc $z(\cdot)$ satisfies

$$-\dot{z}(t) \in G(z(t)) \qquad \text{for a.e. } t \geq 0.$$

Then

$$\|\dot{z}(t)\| = \mathrm{dist}(0, G(z(t))) \qquad \text{a.e.}$$

and therefore

$$\boxed{\varphi(z(0)) - \varphi(z(t)) = \int_0^t \mathrm{dist}^2\left(0; G(z(\tau))\right) d\tau, \qquad \forall t \geq 0.}$$

Problems that Admit Chain Rule

- Convex (Brézis '73, Bruck '75)

Problems that Admit Chain Rule

- Convex (Brézis '73, Bruck '75)

- Subdifferentially regular: any $v \in \partial f(x)$ satisfies

$$f(y) \geq f(x) + \langle v, y - x \rangle + o(\|y - x\|) \quad \text{as } y \to x$$

Problems that Admit Chain Rule

- Convex (Brézis '73, Bruck '75)

- Subdifferentially regular: any $v \in \partial f(x)$ satisfies

$$f(y) \geq f(x) + \langle v, y - x \rangle + o(\|y - x\|) \quad \text{as } y \to x$$

- Whitney stratifiable (D-Drusvyatskiy-Kakade-Lee '18)

Problems that Admit Chain Rule

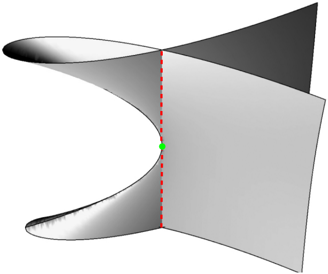
- Convex (Brézis '73, Bruck '75)

- Subdifferentially regular: any $v \in \partial f(x)$ satisfies

$$f(y) \geq f(x) + \langle v, y - x \rangle + o(\|y - x\|) \quad \text{as } y \to x$$

- Whitney stratifiable (D-Drusvyatskiy-Kakade-Lee '18)

  Informally: Graph decomposes into manifolds, fit together in reg. pattern.



Whitney Stratifiable Functions are Ubiquitous

- Virtually exhaustive in optimization.

Whitney Stratifiable Functions are Ubiquitous

- Virtually exhaustive in optimization.
  - **Semianalytic functions**: Any function with graph of the form

$$\bigcup_{i,j=1}^{m} \{x \in \mathbb{R}^d : p_{i,j}(x) \leq 0 \quad \forall i = 1, \ldots, m\}$$

with real-analytic $p_{i,j}$. (Łojasiewicz '65)

Whitney Stratifiable Functions are Ubiquitous

- Virtually exhaustive in optimization.
  - **Semianalytic functions**: Any function with graph of the form

$$\bigcup_{i,j=1}^{m} \{x \in \mathbb{R}^d : p_{i,j}(x) \leq 0 \quad \forall i = 1, \ldots, m\}$$

    with real-analytic $p_{i,j}$. (Łojasiewicz '65)

  - **Definable functions.** Any function with graph definable in o-minimal structure
    - Polynomials, $x^{1/r}$, $\lambda(X)$, $\max\{0, t\}$, $\log(1 + e^t)$, and their sums, products, compositions are definable.
    - Any deep network built from definable pieces.

    (van den Dries-Miller '96,Ta Lê Loi '97)

Whitney Stratifiable Functions Admit Chain Rule

Intuition:

Whitney Stratifiable Functions Admit Chain Rule

Intuition:

- **Stratify.** Domain of φ stratifies into manifolds $M_1, \ldots, M_n$ such that

$$\varphi\Big|_{M_i} \text{ is smooth.}$$

Whitney Stratifiable Functions Admit Chain Rule

Intuition:

- **Stratify.** Domain of φ stratifies into manifolds $M_1, \ldots, M_n$ such that

$$\varphi\Big|_{M_i} \text{ is smooth.}$$

- **Local Chain Rule.** On each manifold $\varphi\Big|_{M_i}$ admits chain rule.

Whitney Stratifiable Functions Admit Chain Rule

Intuition:

- **Stratify.** Domain of φ stratifies into manifolds $M_1, \ldots, M_n$ such that
$$\varphi\Big|_{M_i} \text{ is smooth.}$$
- **Local Chain Rule.** On each manifold $\varphi\Big|_{M_i}$ admits chain rule.
- **Glue Along Arc.** Glue all chain rules along arc $-\dot{z} \in G(z)$ using "Whitney condition" and projection formula of (Daniilidis-Bolte-Lewis-Shiota '07).

Main Result

Thm: (D-Drusvyatskiy-Kakade-Lee '18)

For a stratifiable problem, a.s. all limit points x^* of stochastic proximal subgradient iterates $\{x_k\}$ are critical.

Main Result

Thm: (D-Drusvyatskiy-Kakade-Lee '18)

For a stratifiable problem, a.s. all limit points x^* of stochastic proximal subgradient iterates $\{x_k\}$ are critical.

- Weak Sard from (Bolte-Daniilidis-Lewis-Shiota '07).

Main Result

Thm: (D-Drusvyatskiy-Kakade-Lee '18)

For a stratifiable problem, a.s. all limit points x^* of stochastic proximal subgradient iterates $\{x_k\}$ are critical.

- Weak Sard from (Bolte-Daniilidis-Lewis-Shiota '07).
- Similar result and technique apply to "discretizations" of $-\dot{z} \in G(z)$.

Main Result

Thm: (D-Drusvyatskiy-Kakade-Lee '18)

For a stratifiable problem, a.s. all limit points x^* of stochastic proximal subgradient iterates $\{x_k\}$ are critical.

- Weak Sard from (Bolte-Daniilidis-Lewis-Shiota '07).
- Similar result and technique apply to "discretizations" of $-\dot{z} \in G(z)$.
- Not true for general Lipschitz problems.

Main Result

Thm: (D-Drusvyatskiy-Kakade-Lee '18)

For a stratifiable problem, a.s. all limit points x^* of stochastic proximal subgradient iterates $\{x_k\}$ are critical.

- Weak Sard from (Bolte-Daniilidis-Lewis-Shiota '07).
- Similar result and technique apply to "discretizations" of $-\dot{z} \in G(z)$.
- Not true for general Lipschitz problems.
- **The result is entirely geometric, independent of problem presentation.**

Related Work

- **Weakly Convex.** projected stochastic subgradient (Nurminski '73, '74), stochastic prox-linear (Duchi-Ruan '17).

Related Work

- **Weakly Convex.** projected stochastic subgradient (Nurminski '73, '74), stochastic prox-linear (Duchi-Ruan '17).

- **Semismooth/Generalized Differentiable.** variants of projected stochastic subgradient (Norkin '86 and Ermoliev-Norkin '98)

Related Work

- **Weakly Convex.** projected stochastic subgradient (Nurminski '73, '74), stochastic prox-linear (Duchi-Ruan '17).
- **Semismooth/Generalized Differentiable.** variants of projected stochastic subgradient (Norkin '86 and Ermoliev-Norkin '98)
- **Subdifferentially Regular.** class of stochastic algorithms (Majewski-Miasojedow-Moulines '18) (concurrent with our work)

Related Work

- **Weakly Convex.** projected stochastic subgradient (Nurminski '73, '74), stochastic prox-linear (Duchi-Ruan '17).
- **Semismooth/Generalized Differentiable.** variants of projected stochastic subgradient (Norkin '86 and Ermoliev-Norkin '98)
- **Subdifferentially Regular.** class of stochastic algorithms (Majewski-Miasojedow-Moulines '18) (concurrent with our work)

▶ **Strict descent holds** for all above examples.

Related Work

- **Weakly Convex.** projected stochastic subgradient (Nurminski '73, '74), stochastic prox-linear (Duchi-Ruan '17).
- **Semismooth/Generalized Differentiable.** variants of projected stochastic subgradient (Norkin '86 and Ermoliev-Norkin '98)
- **Subdifferentially Regular.** class of stochastic algorithms (Majewski-Miasojedow-Moulines '18) (concurrent with our work)

▶ **Strict descent holds** for all above examples.

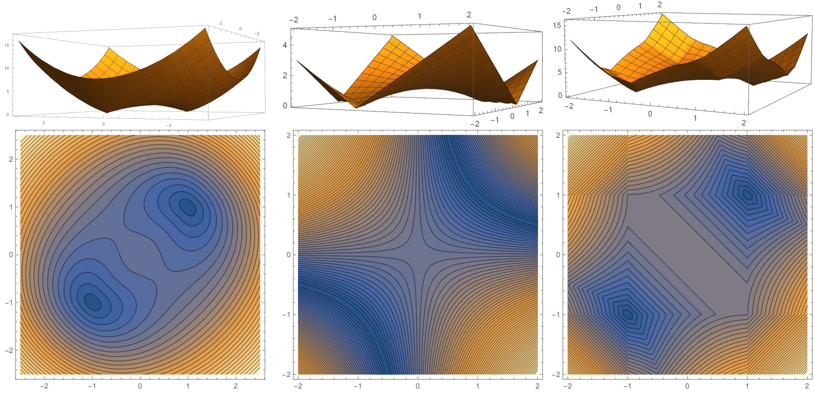
- **Applications of Chain Rule to Deep Learning.**
  - (Du, Hu, Lee '18), (Castera-Bolte-Févotte-Pauwels '19), (Lyu-Li '19)

Broader Perspectives for Nonsmooth Optimization in Data Science

Qualitative Guarantees? Stratifiable Functions.

Quantitative Guarantees? Weakly convex.

Prevalence? Wide.



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

(b) $|xy - 1|$ (blind deconvolution)

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

Quantitative Guarantees and Consequences for Data Science

Weakly Convex

(D-Drusvyatskiy '18)

Sublinear Rates of Moreau Envelope

$$\mathbb{E}\left[\|\nabla \varphi_\lambda(x_{k^*})\|\right] = O(k^{1/4})$$

Key: Moreau (almost) Lyapunov

Quantitative Guarantees and Consequences for Data Science

Weakly Convex

(D-Drusvyatskiy '18)

Sublinear Rates of Moreau Envelope

$$\mathbb{E}\left[\|\nabla \varphi_\lambda(x_{k^*})\|\right] = O(k^{1/4})$$

Key: Moreau (almost) Lyapunov

Weakly Convex + Sharp growth

$$\varphi(x) - \inf_{\mathcal{X}} \varphi \geq \mu \cdot \operatorname{dist}(x, \mathcal{X}^*)$$

(D-Drusvyatskiy '18, '19)

Deterministic/Stochastic Linear Rates

$$\operatorname{dist}(x_k, \mathcal{X}^*) = O\left(\left(1 - \frac{\mu^2}{L^2}\right)^k\right)$$

Key: geometrically decaying stepsize.

Quantitative Guarantees and Consequences for Data Science

Weakly Convex

(D-Drusvyatskiy '18)

Sublinear Rates of Moreau Envelope

$$\mathbb{E}\left[\|\nabla \varphi_\lambda(x_{k^*})\|\right] = O(k^{1/4})$$

Key: Moreau (almost) Lyapunov

Weakly Convex + Sharp growth

$$\varphi(x) - \inf_{\mathcal{X}} \varphi \geq \mu \cdot \mathrm{dist}(x, \mathcal{X}^*)$$

(D-Drusvyatskiy '18, '19)

Deterministic/Stochastic Linear Rates

$$\mathrm{dist}(x_k, \mathcal{X}^*) = O\left(\left(1 - \frac{\mu^2}{L^2}\right)^k\right)$$

Key: geometrically decaying stepsize.

In Practice

Robust phase retrieval, blind deconvolution, low-rank matrix recovery (Eldar-Mendelson '12) (Duchi-Ruan '18) (Charisoplous-D-Diaz-Drusvyatskiy '19) (Li, Zhu, So, Vidal '19)

Weakly Convex + Sharp Growth w.h.p.

Optimal guarantees with out-of-the-box subgradient method

Ex. Phase Retrieval & Blind Deconv

cost $O(md)$.

Summary

- Generic procedure for analyzing discretization schemes of $-\dot{z} \in G(z)$, including stochastic proximal subgradient algorithm.
- Convergence reduced to checking natural property of loss function.
- Identified strict descent, Sard property, and chain rule as key to convergence—automatic for stratifiable losses.

Thanks!