Nonsmooth and nonconvex optimization under statistical assumptions

Damek Davis[1]

School of Operations Research and Information Engineering
Cornell University

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

Smooth nonconvex optimization under statistical assumptions

Empirical risk minimization.

$$\min_x \; f(x; A)$$

Hard in general, but "easy" when A is random.

Example: Matrix Completion.

- Observe a random subset of entries $A \subseteq [n]^2$ of a low rank matrix M.
- Find M by optimizing

$$f(L, R; A) = \frac{1}{2}\|\Pi_A(LR^T - M)\|^2$$

- With appropriate regularization, all local minimizers are global minimizers.[2]

[2] Ge, Lee, Ma. Matrix Completion has No Spurious Local Minimum (2016)

Smooth nonconvex optimization under statistical assumptions

Further Examples. Provable complexity guarantees for Matrix Completion/Sensing, Tensor Recovery/Decomposition and Latent Variable Models, Phase retrieval, Dictionary Learning, Deep Learning, Nonnegative/Sparse Principal Component Analysis, Mixture of Linear Regression, Super Resolution, Synchronization and Community Detection, Joint Alignment Problems, and System Identification.

Extensive list. http://sunju.org/research/nonconvex/

Smooth nonconvex optimization under statistical assumptions

Coarsest approach.

1. Find initial solution estimate $\hat{x}$.
   - Typically found via spectral method (min/max eigenvector).
2. Run a "local search method."
   - Very often gradient descent.

Smooth nonconvex optimization under statistical assumptions

Fine-grained approach.

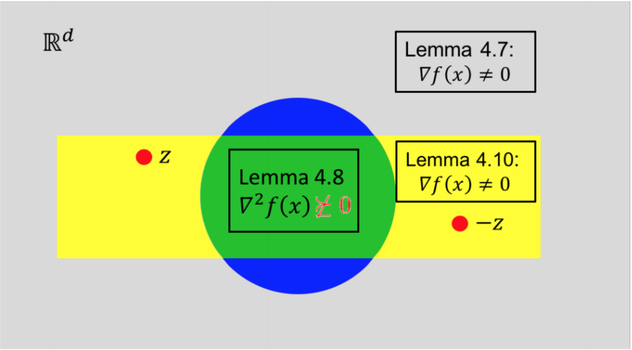
1. Characterize geometry of loss function:
   - Large gradient region
   - Negative curvature region
   - Local strong convexity around minimizers

2. Gradient descent with random initialization converges to minimizers.[3]

[3] Lee, Simchowitz, Jordan, Recht. Gradient Descent Converges to Minimizers. (2016)

Smooth nonconvex optimization under statistical assumptions[4]

objective $f(x) = \|P_\Omega(M) - P_\Omega(xx^\top)\|_F^2 + \lambda R(x)$



 $\left\{x \colon \|x\|_\infty \leq \frac{4\mu}{\sqrt{d}}\right\}$ ● local (and global) min

 $\left\{x \colon \|x\|^2 \leq \frac{1}{16}\right\}$

[4] Ge, Lee, Ma. Matrix Completion has No Spurious Local Minimum (2016)

Smooth nonconvex optimization under statistical assumptions

How to Characterize Geometry.

1. Analyze **population** risk

$$\mathbb{E}_A \left[f(x; A) \right]$$

   Randomness "integrated out." Typically simple function.

2. "Transfer" geometry of population model back to **empirical** risk

$$f(x; A),$$

   using concentration inequalities.

   - Gradients and Hessians of empirical risk often concentrate around population gradients and Hessians.

Smooth nonconvex optimization under statistical assumptions

General framework for smooth geometry transfer.[5]

Assume that gradients are subgaussian random variables:

$$\mathbb{E}_A\left[\exp\left(\langle v, \nabla f(x,A) - \mathbb{E}_A\left[\nabla f(x,A)\right]\rangle\right)\right] \le \exp\left(\frac{\tau^2\|v\|^2}{2}\right) \quad \forall v \in \mathbb{R}^d$$

Union bound leads to "optimal" concentration:

$$\mathbb{P}\left(\sup_{x\in B}\|\nabla f(x,A) - \mathbb{E}_A\left[\nabla f(x,A)\right]\| \le \tau^2 \cdot \sqrt{\frac{c\log(1/\delta)d\log n}{n}}\right) \ge 1-\delta$$

where n is the number of "measurements."

Similar results hold for Hessians as well.

[5] Mei, Yu, Montanari. The landscape of empirical risk for non-convex losses (2016)

Smooth nonconvex optimization under statistical assumptions

Conclusions.

- The pipeline is well-understood.
- Techniques typically tailored to individual problems.

What to do in the nonsmooth setting?

What to do in the nonsmooth setting?

Why should we care?

1. ℓ_1-type losses insensitive to outliers/enforce sparsity.
2. ReLU ($\max\{0, x\}$) nonsmooth activation units in deep networks very successful in practice.
3. Even in traditional nonlinear programming, difficult constraints $c(x) = 0$, typically enforced with exact penalty:

$$\|c(x)\|.$$

Nonsmooth nonconvex optimization under statistical assumptions

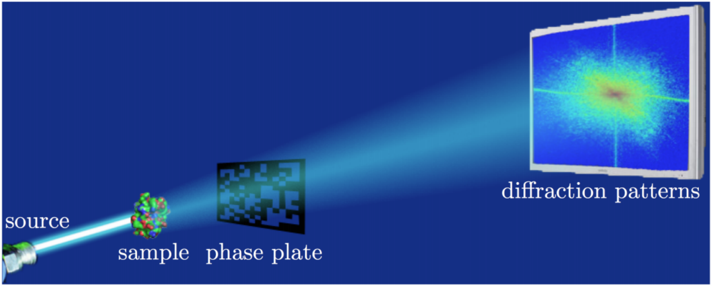
What fails for nonsmooth?

1. Unclear what "local-search" should mean.

2. Geometry
   - No good quantifiable concept of saddle points (negative eigenvalue of Hessian).
   - Strong convexity $\not\Longrightarrow$ fast convergence.
   - Subdifferentials do not concentrate.

What's coming. Develop general principles for nonsmooth setting, guided by concrete application.

Phase Retrieval[6]



[6] Candes, Li, Soltanolkotabi. Phase Retrieval from Coded Diffraction Patterns (2013)

Example: nonsmooth phase retrieval

"Real" Phase Retrieval.

1. Given signal $\bar{x} \in \mathbb{R}^d$.
2. We observe squared magnitude of dot product

$$\boxed{b_i = \langle a_i, \bar{x} \rangle^2 \qquad i = 1, \ldots, n}$$

with several measurement vectors a_i.

- NP-Hard in worst case.[7]
- Becomes "easy" with subgaussian and "well-spread" a_i.
- Only solutions: $\{\pm \bar{x}\}$ if $n = \Omega(d)$.

[7] Fickus, Mixon, Nelson, Yang. Phase retrieval from very few measurements (2014)

Example: nonsmooth phase retrieval

Empirical Risk.

$$
\begin{aligned}
f_E(x) :=& \frac{1}{n}\sum_{i=1}^{n} |\langle a_i, x\rangle^2 - \langle a_i, \bar{x}\rangle^2| \\
=& \frac{1}{n}\|(Ax)^2 - b\|_1.
\end{aligned}
$$

- Nonsmooth and nonconvex.
- If $n = \Omega(d)$, minimizers $\pm\bar{x}$.
- **"Robust:"** can corrupt $\approx 1/2$ of $\langle a_i, \bar{x}\rangle^2$ in arbitrary way.
  - Key is nonsmooth formulation
  - Lose robustness with smooth formulations.

Example: nonsmooth phase retrieval

Key Questions.

1. Linearly convergent algorithm?

2. Stationary point structure?

Linearly convergent algorithm for nonsmooth nonconvex?

Fast local convergence requires "regularity."

- In smooth case, "regularity" = local strong convexity.
- In nonsmooth case "regularity" = μ-**sharpness:**

$$f(x) - \inf f \geq \mu \cdot \underbrace{\operatorname{dist}(x, \arg\min f)}_{\text{distance to solution set}}$$

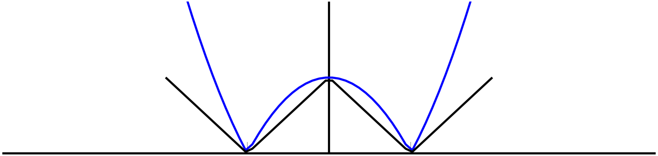


Figure: $f(x) = |x^2 - 1|$ (blue) and $\operatorname{dist}(x; \{\pm 1\})$ (black).

Sharpness of f_E

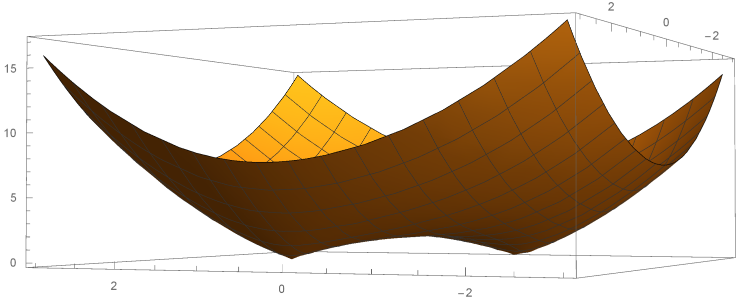
Theorem (Eldar-Mendelson (2012))

f_E is $\Omega(\|\bar{x}\|)$ sharp.

Proved that

$$f_E(x) - \inf f_E \geq \kappa \cdot \|x - \bar{x}\| \|x + \bar{x}\|$$

"Strong Stability"



Interlude: convexity + sharpness

Consider convex minimization problem:

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

- f is Lipschitz and μ-**sharp.**

Polyak subgradient method:

$$v_k \in \partial f(x_k)$$

$$x_{k+1} = x_k - \boxed{\frac{f(x_k) - \inf f}{\|v_k\|^2}} \cdot v_k$$

- Linearly converges (Polyak 1969).

Adapt Polyak method to nonconvex setting?

Weak convexity:

$$f + \frac{\rho}{2}\|\cdot\|^2 \qquad \text{is convex,}$$

where $\rho > 0$.

Weakly convex class is broad.

- Convex Composite: Includes all functions

  $$h \circ c$$

  where h is convex and Lipschitz and c is a smooth map.

Example: Convex Composite

1. **Robust PCA.** Given $\bar{M} = \bar{L}\bar{R}^T + \bar{S} \in \mathbb{R}^{m\times n}$ (low rank + sparse)

$$f(L,R) = \frac{1}{nm}\|LR^T - \bar{M}\|_1$$

$\implies f + \|\cdot\|^2$ is convex

Example: Convex Composite

1. **Robust PCA.** Given $\bar{M} = \bar{L}\bar{R}^T + \bar{S} \in \mathbb{R}^{m \times n}$ (low rank + sparse)

$$f(L, R) = \frac{1}{nm} \|LR^T - \bar{M}\|_1$$

$\implies f + \|\cdot\|^2$ is convex

2. **Phase Retrieval.** $f_E + 5\|\cdot\|^2$ is convex (w.h.p if $a_i \sim \mathsf{N}(0, I_d)$)

Not weakly convex

1. **Negative ℓ_1.** $f(x) = -\|x\|_1$

Not weakly convex

1. **Negative ℓ_1.** $f(x) = -\|x\|_1$

2. **Canonical robust phase retrieval.** Given $b_i = |\langle a_i, \bar{x} \rangle|$

$$f(x) = \frac{1}{m} \sum ||\langle a_i, x \rangle| - b_i|$$

Not weakly convex

1. **Negative ℓ_1.** $f(x) = -\|x\|_1$

2. **Canonical robust phase retrieval.** Given $b_i = |\langle a_i, \bar{x} \rangle|$

$$f(x) = \frac{1}{m} \sum ||\langle a_i, x \rangle| - b_i|$$

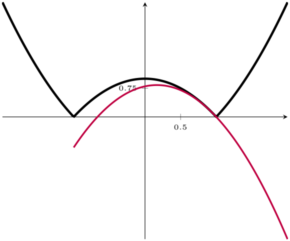
3. **Neural networks.** Simple neural network (with data (x_j, b_j))

$$f(w) = \frac{1}{2n} \sum_{j=1}^{n} \left(\sum_{i=1}^{k} \max\{w_i^T x_j, 0\} - b_j \right)^2$$

Subgradients for weakly convex

Natural subdifferential: $v \in \partial f(x) \iff$

$$f(y) \geq f(x) + \langle v, y - x\rangle - \frac{\rho}{2}\|y - x\|^2 \quad \forall y.$$



$$f(x) = |x^2 - 1|$$

Stationary points of sharp + weakly convex

Lemma (D., Drusvyatskiy, Paquette (2017))

If f is ρ-weakly convex and μ-sharp, then the tube

$$\mathcal{T} := \left\{ x \mid \operatorname{dist}(x, \arg\min f) < \frac{2\mu}{\rho} \right\}$$

contains no stationary points.

- Denote $\mathcal{S} := \arg\min f$.
- Chose stationary $x \notin \mathcal{S}$: $0 \in \partial f(x)$
- Choose $\bar{x} \in \mathcal{S}$ so that $\|x - \bar{x}\| = \operatorname{dist}(x, \mathcal{S})$.

Stationary points of sharp + weakly convex

Lemma (D., Drusvyatskiy, Paquette (2017))

If f is ρ-weakly convex and μ-sharp, then the tube

$$\mathcal{T} := \left\{ x \mid \operatorname{dist}(x, \arg\min f) < \frac{2\mu}{\rho} \right\}$$

contains no stationary points.

- Denote $\mathcal{S} := \arg\min f$.
- Chose stationary $x \notin \mathcal{S}$: $0 \in \partial f(x)$
- Choose $\bar{x} \in \mathcal{S}$ so that $\|x - \bar{x}\| = \operatorname{dist}(x, \mathcal{S})$.

$$\mu \cdot \operatorname{dist}(x, \mathcal{S}) \underbrace{\leq}_{\text{sharpness}} f(x) - f(\bar{x}) \underbrace{\leq}_{\text{weak convexity}} \frac{\rho}{2}\|x - \bar{x}\|^2 = \frac{\rho}{2}\operatorname{dist}^2(x, \mathcal{S})$$

Therefore,

$$\frac{2\mu}{\rho} \leq \operatorname{dist}(x, \mathcal{S}).$$

Polyak for sharp + weakly convex

Theorem (D., Drusyatskiy, Paquette (2017))

Polyak method linearly converges when initialized in $\mathcal{T}$.

- Follow up work for case when $\inf f$ is not known.[8]
- Little was known about convergence rates of subgradient methods for nonconvex problems until quite recently.[9] [10]
- Other problems
  - Covariance estimation, blind deconvolution, robust PCA, matrix completion....[11]

---

[8] D., Drusvyatskiy, MacPhee, Paquette (2018)

[9] D., Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems (2017)

[10] D., Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. (2018)

[11] Charisopoulos, Chen, D., Diaz, Ding, Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. (2019)

Consequences for phase retrieval

Theorem (D., Drusvyatskiy, Paquette (2017))

Suppose $n = \Omega(d)$. After spectral initialization, the Polyak method converges linearly on f_E.

- In phase retrieval, $\mu = \Omega(\|\bar{x}\|)$, $\rho = O(1)$

$$\mathcal{T} = \left\{ x \mid \frac{\operatorname{dist}(x, \{\pm\bar{x}\})}{\|\bar{x}\|} = O(1) \right\}.$$

- Spectral initialization can produce initializer in $\mathcal{T}$.[12]
- Cost per iteration is two matrix multiplications

$$\frac{2}{n} \sum_{i=1}^{n} \langle a_i, x\rangle \operatorname{sign}(\langle a_i, x\rangle^2 - \langle a_i, \bar{x}\rangle^2) a_i \in \partial f_E(x).$$

[12] Duchi, Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. (2017)

Polyak for sharp + weakly convex: experiment

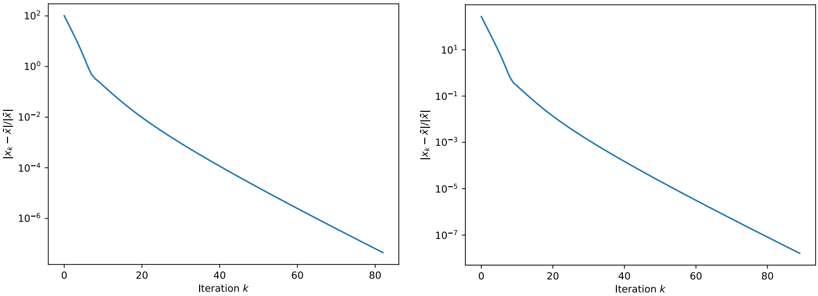


Figure: Convergence plot on two different images taken from the Hubble telescope (iterates vs. $\|x_k - \bar{x}\|/\|\bar{x}\|$). The dimensions of the problem on the left are $d \approx 2^{22}$ and $m = 3d \approx 2^{24}$. The dimensions of the problem on the right are $d \approx 2^{24}$ and $m = 3d \approx 2^{25}$. For the plot on the left, the entire experiment, including initialization and the subgradient method completed in 3 min. For the plot on the right, it completed in 25.6 min. The majority of time ≈ 25 min was taken up by the initialization. The results were obtained on a standard desktop: Intel(R) Core(TM) i7-4770 CPU3.40 GHz with 8.00 GB RAM.

Comparison to smooth case

$$f_S(x) = \frac{1}{n}\sum_{i=1}^{n} |\langle a_i, x\rangle^2 - \langle a_i, \bar{x}\rangle^2|^2$$

- **Poorly conditioned** near $\{\pm\bar{x}\}$:

$$\frac{1}{2}I \preceq \nabla^2 f_S(x) \preceq O(d)I.$$

- **Overly pessimistic** contraction factor:

$$\|x_{k+1} - \bar{x}\| \leq (1 - O(1/d))\|x_k - \bar{x}\|.$$

  To overcome, carefully analyze trajectory of gradient descent.[13]

- **Nonsmooth** Polyak fast "out-of-the-box:" constant contraction factor.

[13] Chi, Lu, and Chen. Nonconvex Optimization Meets Low-Rank Matrix Factorization: An Overview (2019)

Example: nonsmooth phase retrieval

Key Questions.

1. Linearly convergent algorithm?

2. Stationary point structure?

Population model

Population model (Gaussian case):

$$f_P(x) := \mathbb{E}_{a\sim \mathsf{N}(0,I_d)}\left[\left|\langle a,x\rangle^2 - \langle a,\bar{x}\rangle^2\right|\right]$$

Explicit form: with $X := xx^T - \bar{x}\bar{x}^T$

$$f_P(x) = \frac{4}{\pi}\left[\mathrm{Tr}(X)\cdot \arctan\left(\sqrt{\left|\frac{\lambda_{\max}(X)}{\lambda_{\min}(X)}\right|}\right) + \sqrt{|\lambda_{\max}(X)\lambda_{\min}(X)|}\right] - \mathrm{Tr}(X).$$

- How to characterize stationary points of f_P?

Spectral function characterization

Lemma (D., Drusvyatskiy, Paquette (2017))

There is a symmetric convex function g_P satisfying

$$f_P(x) = g_P(\lambda(X)).$$

where $\lambda(X)$ is the vector of eigenvalues of $X := xx^T - \bar{x}\bar{x}^T$.

- Still nonconvex and nonsmooth.
- Exploit symmetries to characterize stationary points.
- Instead of thinking about f_P, analyze all functions of the same form.

Subgradients of spectral functions

Consider

$$f(x) := g(\lambda(xx^T - \bar{x}\bar{x}^T)) \qquad g \text{ finite, symmetric, convex}$$

Chain rule shows that

$$\partial f(x) = 2\partial(g \circ \lambda)(X)x$$

Transfer Principle (Lewis 1999).

$$V \in \partial(g \circ \lambda)(X)$$

$$\Updownarrow$$

there is an orthogonal matrix U satisfying

1. $\lambda(V) \in \partial g(\lambda(X))$
2. $V = U\mathrm{diag}(\lambda(V))U^T$
3. $X = U\mathrm{diag}(\lambda(X))U^T$

Stationary points of spectral functions

Theorem (D., Drusvyatskiy, Paquette)

Suppose that x is stationary for f, that is $Vx = 0$. Then one of the following conditions holds:

1. $f(x) \leq f(\bar{x})$
2. $x = 0$
3. $\langle x, \bar{x} \rangle = 0$, $\lambda_1(V) = 0$.

Moreover, if $\bar{x}$ minimizes f, then a point x is stationary for f if and only if x satisfies 1, 2, or 3.

- Point $\bar{x}$ minimizes f_P.
- $\implies$ Nontrivial stationary points of f_P determined by $\lambda_1(V) = 0$.

Stationary points of population model

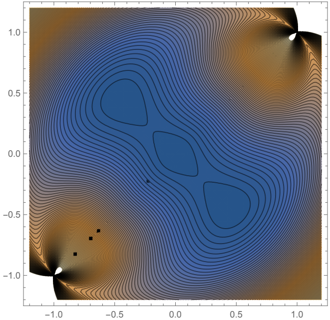
Theorem (D., Drusvyatskiy, Paquette)

The stationary points of the population objective f_P are precisely

$$\{0\} \cup \{\pm \bar{x}\} \cup \{x \in \bar{x}^{\perp} : \|x\| = c \cdot \|\bar{x}\|\},$$

where $c > 0$ (approx. $c \approx 0.4416$) is the unique solution of the equation $\frac{\pi}{4} = \frac{c}{1+c^2} + \arctan(c)$.

Gradient: $x \mapsto \|\nabla f_P(x)\|$.



Stationary points of empirical risk?

$\partial f_E(x)$ can be poor pointwise approximation of $\partial f_P(x)$.

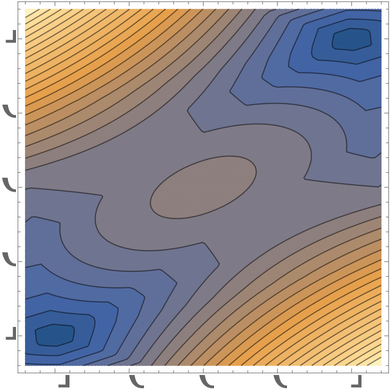


Figure: Level sets of f_E

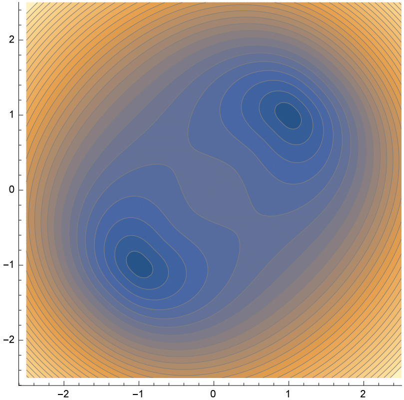


Figure: Level sets of f_P

Function value concentration

Theorem (Eldar-Mendelson (2012))

With high probability

$$|f_E(x) - f_P(x)| \leq C \cdot \sqrt{\frac{d}{n}} \|x - \bar{x}\| \|x + \bar{x}\| \qquad \textit{for all } x \in \mathbb{R}^d.$$

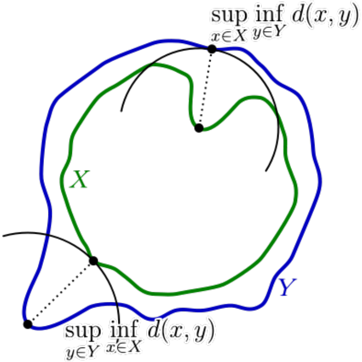
Does function value approximation imply any "closeness" of subdifferentials?

- Hausdorff distance plays key role.

Hausdorff distance

The **Hausdorff distance** between sets X and Y:

$$\mathrm{dist}_H(X,Y) = \max\{\sup_{x\in X} \mathrm{dist}(x,Y), \sup_{y\in Y} \mathrm{dist}(y,X)\}.$$



Closeness of subdifferential graphs

Define **graph** of subdifferential of function f:

$$\text{gph } \partial f = \{(x, v) \mid v \in \partial f(x)\} \subseteq \mathbb{R}^{d \times d}.$$

[14] Attouch, Wets. Quantitative stability of variational systems: the epigraphical distance. (1989)

Closeness of subdifferential graphs

Define **graph** of subdifferential of function f:

$$\text{gph } \partial f = \{(x, v) \mid v \in \partial f(x)\} \subseteq \mathbb{R}^{d \times d}.$$

Theorem (D., Drusvyatskiy, Paquette (2017))

Given two ρ-weakly convex functions f and g satisfying

$$|f(x) - g(x)| \leq \delta,$$

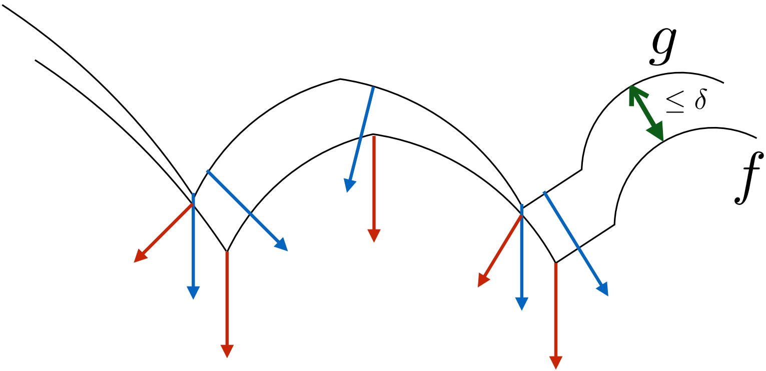
the bound holds:

$$\text{dist}_H(\text{gph } \partial f, \text{gph } \partial g) \leq \sqrt{4(\rho + \sqrt{2 + \rho^2})} \cdot \sqrt{2\delta} = O(\sqrt{\delta})$$

- Can adapt to non constant $\delta(x)$ (“small function”).
- Quantitative version of Attouch-Wets' variational principle.[14]

[14] Attouch, Wets. Quantitative stability of variational systems: the epigraphical distance. (1989)

Closeness of subdifferential graphs



Stationary points of empirical risk

Apply previous result to locate stationary points.

Theorem (D., Drusvyatskiy, Paquette (2017))

Every stationary point of f_E satisfies $\|x\| \lesssim \|\bar{x}\|$ and one of the two conditions:

$$\frac{\|x\|\|x-\bar{x}\|\|x+\bar{x}\|}{\|\bar{x}\|^3} \lesssim \sqrt[4]{\frac{d}{m}} \qquad or \qquad \left\{ \begin{array}{c} \left|\frac{\|x\|}{\|\bar{x}\|} - c\right| \lesssim \sqrt[4]{\frac{d}{m}} \cdot \left(1 + \frac{\|\bar{x}\|}{\|x\|}\right) \\ \frac{|\langle x, \bar{x}\rangle|}{\|x\|\|\bar{x}\|} \lesssim \sqrt[4]{\frac{d}{m}} \cdot \frac{\|\bar{x}\|}{\|x\|} \end{array} \right\},$$

where $c > 0$ is the unique solution of the equation $\frac{\pi}{4} = \frac{c}{1+c^2} + \arctan(c)$.

- Compare to stationary points of $\partial f_P(x)$.

$$\{0\} \cup \{\pm\bar{x}\} \cup \{x \in \bar{x}^{\perp} : \|x\| = c \cdot \|\bar{x}\|\},$$

Extensions of ideas

Phase retrieval was vehicle to understand nonsmooth setting.

- **Recovery Problems.** Covariance estimation, blind deconvolution, matrix completion, robust PCA formulations are **sharp** and **weakly convex**....[15]

- **Concentration for subdifferentials graphs.**[16] Statistical learning (ERM/SAA) with weakly convex losses:

$$f_P(x) := \mathbb{E}_z\left[f(x;z)\right] \qquad\qquad f_E(x) := \frac{1}{n}\sum_{i=1}^{n} f(x;z_i)$$

$$\Longrightarrow \boxed{\mathrm{dist}_H(\mathrm{gph}\ \partial f_P, \mathrm{gph}\ \partial f_E) = \widetilde{O}(\sqrt{L^2 d/n})}$$

- **Algorithms.** Toolbox for large-scale nonsmooth nonconvex problems.

---

[15] Charisopoulos, Chen, D., Diaz, Ding, Drusvyatskiy. Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. (2019)

[16] D. and Drusvyatskiy Graphical Convergence of Subgradients in Nonconvex Optimization and Learning. (2018)

Subgradient methods for nonsmooth nonconvex optimization

- **Open problem solved:** complexity of stochastic proximal subgradient method for weakly convex problems.[17] Further analyzed any "model-based" algorithm.[18] New idea: use smooth potential function for nonsmooth problems.

- Linearly convergent subgradient methods without optimal value.[19] Similar techniques as in convex setting.

- **Open problem solved:** Proved stochastic subgradient method converges to stationary points for **virtually exhaustive class** of nonpathological (including all semialgebraic) functions.[20] Convergence was not known beyond weakly convex problems. New idea: such functions have well-behaved differential inclusions $\dot{z}(t) \in -\partial f(z(t))$.

[17] D. and Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions (2018)

[18] D. and Drusvyatskiy. Stochastic model-based minimization of weakly convex functions (2018)

[19] D. and Drusvyatskiy, MacPhee, and Paquette. Subgradient methods for sharp weakly convex functions (2018)

[20] D., Drusvyatskiy, Kakade, Lee. Stochastic subgradient method converges on tame functions (2018)

Thanks!

- The nonsmooth landscape of phase retrieval. (2017)
  D., Drusvyatskiy, Paquette. IMA Journal of Numerical Analysis
- Subgradient methods for sharp weakly convex functions. (2018)
  D., Drusvyatskiy, MacPhee, Paquette. JOTA
- Stochastic model-based minimization of weakly convex functions. (2018)
  D., Drusvyatskiy. SIOPT
- Stochastic subgradient method converges on tame functions. (2018)
  D., Drusvyatskiy, Kakade, Lee. FOCM
- Graphical Convergence of Subgradients in Nonconvex Optimization and Learning. (2018)
  D., Drusvyatskiy, Kakade, Lee. arXiv:1810.07590
- Composite optimization for robust blind deconvolution. (2019)
  Charisopoulos, D., Díaz, Drusvyatskiy. arXiv:1901.01624
- Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence. (2019)
  Charisopoulos, Chen, D., Ding, Díaz, Drusvyatskiy. arXiv:1904.10020