Run and Inspect Method: Global Bounds for R-Local Minimizers

Yifan Chen, Yuejiao Sun, Wotao Yin

September 2017

Overview

- **goal**: an approximate global solution for non-convex optimization
- **approach**: run an existing algorithm, inspect its limit
- for high-dimensional problems, introduce new block-wise methods to reduce inspection costs
- **guarantee:** global bounds for convex+"nice nonconvex" objective

Inspection step

- Run an algorithm till its near convergence
- Inspect the R-radius of latest $\mathbf{x}^k$, looking for a sufficient descent point by sampling
  - If found, then resume your algorithm from the point;
  - otherwise, $\mathbf{x}^k$ must be an approx R-local minimizer.
- We will develop a global bound for an approx R-local minimizer

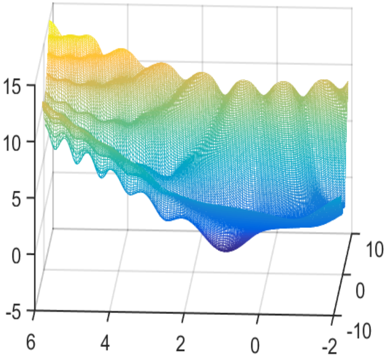
Block coordinate inspection in high dimensions

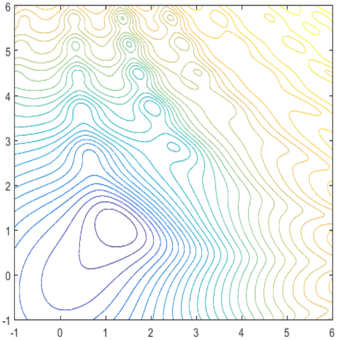
- **curse of dim:** #sample-points is exponential in dimension
- **solution**: decompose to blocks of small dimensions
- Run a (block) coordinate algorithm, inspection each block;
  #sample-points grows linearly in #blocks
- **updated guarantee**: global bounds worsens only linearly in #blocks
  Avoided the curse of dimensionality!

2D numerical experiment

Rugged landscape

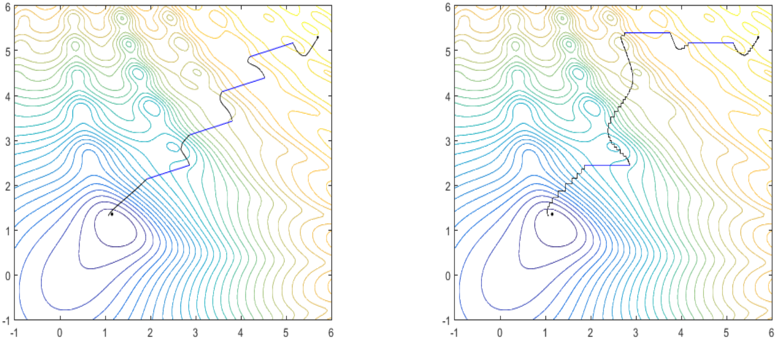
$$F(x,y) = -20\exp\left(-0.04(x^2+y^2)\right) - \exp\left(0.7(\sin(xy)+\sin y)+0.2\sin(x^2)\right) + 20$$





Black line represents the gradient descent step

Blue line represents the inspection step



Iteration trace

Left : Gradient descent(stepsize=1/40) + Global search, $R = 1$

Right : Block gradient descent(stepsize=1/40) + Blockwise search, $\mathbf{R} = (1, 1)$

R-local minimizer

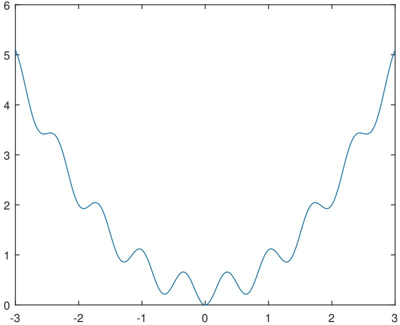
definition: $\bar{\mathbf{x}}$ is an R-local minimizer of function F if

$$F(\bar{\mathbf{x}}) = \min_{\mathbf{x} \in B(\bar{\mathbf{x}}, R)} F(\mathbf{x})$$

- $R = \infty \;\Rightarrow\; \bar{\mathbf{x}}$ is a global minimizer
- $R > 0$ exists $\;\Rightarrow\; \bar{\mathbf{x}}$ a local minimizer
- For a fixed $R > 0$, R-local is between global and local

R-local is global: 1D example

$$F(x) = \frac{x^2}{2} + a\sin\left(b\pi(x - \frac{1}{2b})\right) + a, \text{ where } a = 0.3,\ b = 3$$

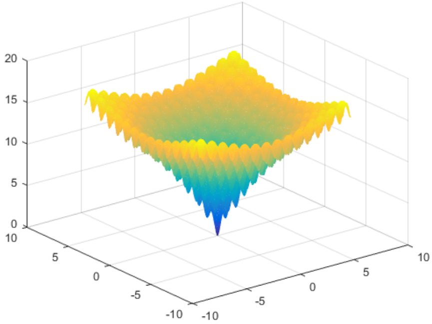


if $R > 2\sqrt{a}$, then 0 is the only R-local minimizer

2D example: Ackley's function

$$F(x,y) = -20e^{-0.2\sqrt{0.5(x^2+y^2)}} - e^{0.5(\cos 2\pi x + \cos 2\pi y)} + e + 20$$

often used to evaluate evolutionary algorithms



for suff. large R, $(0,0)$ is the only R-local minimizer

Theory of (approximate) R-local minimizer

assumptions:

$$F(\mathbf{x}) = f(\mathbf{x}) + r(\mathbf{x})$$

(decomposition is only needed for theoretical analysis)

- f is differentiable and ∇f is L-Lipschitz continuous
- r is "nice": exist $\alpha, \beta \geq 0$ such that

$$|r(\mathbf{x}) - r(\mathbf{y})| \leq \alpha \|\mathbf{x} - \mathbf{y}\| + 2\beta, \quad \forall \mathbf{x}, \mathbf{y}$$

$\|\nabla f\|$ bounds at an (approximate) R-local minimizer:

- If $\bar{\mathbf{x}}$ is an R-local minimizer of F, then

$$\|\nabla f(\bar{\mathbf{x}})\| \leq \alpha + \max\{\frac{4\beta}{R}, 2\sqrt{\beta L}\}$$

  When $R > 2\sqrt{\frac{\beta}{L}}$ the bound will not improve

- If $F(\bar{\mathbf{x}}) \leq \min_{\mathbf{x} \in B(\bar{\mathbf{x}}, R)} F(\mathbf{x}) + \eta$, then

$$\|\nabla f(\bar{\mathbf{x}})\| \leq \alpha + \max\{\frac{4\beta + 2\eta}{R}, \sqrt{(4\beta + 2\eta)L}\}$$

Previous slide establishes $\|\nabla f(\bar{\mathbf{x}})\| \leq \delta$

Now, assume the Polyak-Łojasoewicz inequality (slightly weaker than strong convexity)

$$\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu \left(f(\mathbf{x}) - f^* \right), \quad \forall \mathbf{x}$$

Example satisfying this inequality:

- Strongly convex
- Strongly convex composed with linear $f(\mathbf{x}) = g(A\mathbf{x})$

$$f(\mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2$$

$$f(\mathbf{x}) = \sum_{i=1}^{n} \log(1 + \exp(b_i \mathbf{a}_i^T \mathbf{x})) \quad \text{in compact region}$$

Global optimality bounds

Under above assumptions

- If $\alpha = 0$, then

$$F(\bar{\mathbf{x}}) - F^* \leq \frac{\delta^2}{2\mu} + 2\beta$$

- If $\alpha \geq 0$ and any global minimizer $\mathbf{x}, \mathbf{y}$ of f satisfy $\|\mathbf{x} - \mathbf{y}\| \leq M$ then

$$d(\bar{\mathbf{x}}, \textsf{sol set}) \leq \frac{2\delta}{\mu} + M$$

$$F(\bar{\mathbf{x}}) - F^* \leq \frac{\delta^2 + 2\alpha\delta}{\mu} + \alpha M + 2\beta$$

Obtaining an approximate R-local minimizer

- Suppose a descent algorithm (nearly) converges at $\mathbf{x}^k$
- Inspection samples points $\mathbf{y}_1^k, \mathbf{y}_2^k, ..., \mathbf{y}_{m^k}^k \in B(\mathbf{x}^k, R)$:
  - on coarse to finer grids, or
  - uniformly at random, or
  - nonuniformly according to function properties, or
  - by MCMC or Gibbs
- **hit and run**: once finding a point that decreases objective by $\geq \delta$, resume the algorithm there; otherwise, return $\mathbf{x}^k$

Inspection guarantee

- **assume**:
  - sample in $B(\bar{\mathbf{x}}, R)$ at density[1] r
  - function $F(x)$ $\bar{L}$-Lipschitz in $B(\bar{\mathbf{x}}, R)$
- If no sample is found to improve F by δ, then

$$F(\bar{\mathbf{x}}) \leq \min_{\mathbf{x} \in B(\bar{x}, R)} F(\mathbf{x}) + (\bar{L}r + \delta),$$

that is, $\bar{\mathbf{x}}$ is an R-local minimizer up to $\bar{L}r + \delta$

[1] For any $\mathbf{x} \in B(\bar{\mathbf{x}}, R)$, there exists a sampled point $\mathbf{y}$ such that $\|\mathbf{x} - \mathbf{y}\| \leq r$

Partial summary

Abstract algorithm:

- Run an existing descent algorithm to $\mathbf{x}^k$ with prescribed precision
- Inspect samples in $B(\mathbf{x}^k, R)$
  - if δ-descent is found, resume the algorithm there
  - otherwise, stop and return $\mathbf{x}^k$

The algorithm stops finitely with an approximate R-local minimizer.

If the objective is convex$+$"nice nonconvx", then nearly globally optimal.

Blockwise version

- $\mathbf{x}$ has s blocks
- Let $F(x_i, \bar{\mathbf{x}}_{-i})$ denotes $F(\bar{x}_1, ..., x_i, ..., \bar{x}_s)$ as a function of x_i
  $\bar{\mathbf{x}}$ is a blockwise $\mathbf{R}$-local minimizer of F if

$$F(\bar{x}_i, \bar{\mathbf{x}}_{-i}) = \min_{x_i \in B(\bar{x}_i, R_i)} F(x_i, \bar{\mathbf{x}}_{-i}) \ \ \forall 1 \le i \le s$$

- When $\mathbf{R} = \infty$, $\bar{\mathbf{x}}$ is a Nash equilibrium point
- (Under the same assumptions) bounded gradient

$$\|\nabla f(\bar{\mathbf{x}})\| \le \sqrt{s}\left(\alpha + \max\{\frac{4\beta}{\min_i R_i}, 2\sqrt{\beta L}\}\right).$$

Blockwise inspection

- If the descent method in use is a global method,
  then inspect block by block
- If the descent method in use is a block coordinate descent method,
  then integrate inspection into each block

Useful R-local-min variants

select $R, \gamma > 0$

- $\bar{\mathbf{x}}$ is a R-local prox minimizer of F if

$$F(\bar{\mathbf{x}}) = \min_{\mathbf{x} \in B(\bar{\mathbf{x}}, R)} F(\mathbf{x}) + \frac{\gamma}{2} \|\mathbf{x} - \bar{\mathbf{x}}\|^2,$$

  When $R = \infty$ it is called a prox minimizer

- Suppose $F = F_1 + F_2$;

  $\bar{\mathbf{x}}$ is a R-local prox-linear minimizer of $F = F_1 + F_2$ if

$$F(\bar{\mathbf{x}}) = \min_{\mathbf{x} \in B(\bar{\mathbf{x}}, R)} \langle \nabla F_1(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle + \frac{\gamma}{2} \|\mathbf{x} - \bar{\mathbf{x}}\|^2 + F_2(\mathbf{x}).$$

  When $R = \infty$ it is called a prox-linear minimizer

Blockwise version:

select $\mathbf{R} = (\gamma_1, ..., \gamma_s) > 0$

- $\bar{\mathbf{x}}$ is a blockwise $\mathbf{R}$-local prox minimizer of F if

$$F(\bar{x}_i, \bar{\mathbf{x}}_{-i}) = \min_{x_i \in B(\bar{x}_i, R_i)} F(x_i, \bar{\mathbf{x}}_{-i}) + \frac{\gamma_i}{2}\|x_i - \bar{x}_i\|^2, \quad \forall 1 \le i \le s$$

- Suppose $F = F_1 + F_2$;
  $\bar{\mathbf{x}}$ is a blockwise $\mathbf{R}$-local prox-linear minimizer of F if for $1 \le i \le s$

$$F(\bar{x}_i, \bar{\mathbf{x}}_{-i}) = \min_{x_i \in B(\bar{x}_i, R_i)} \langle \nabla_i F_1(\bar{x}_i, \bar{\mathbf{x}}_{-i}), x_i - \bar{x}_i \rangle + \frac{\hat{\gamma}_i}{2}\|x_i - \bar{x}_i\|^2 + F_2(x_i, \bar{\mathbf{x}}_{-i})$$

Blockwise updating rule

Each step choose a index i_k to update; $\mathbf{R}^k$ converges to $\mathbf{R}$; γ^k converges to γ

1. Blockwise $\mathbf{R}$-local minimization:

$$x_{i_k}^{k+1} \in \operatorname*{arg\,min}_{x_{i_k} \in B(x_{i_k}^k, R_{i_k}^k)} F(x_{i_k}, \mathbf{x}_{-i_k}^k)$$

   Greedy choice of index is needed when $s > 2$

2. Blockwise $\mathbf{R}$-local prox minimization:

$$x_{i_k}^{k+1} \in \operatorname*{arg\,min}_{x_{i_k} \in B(x_{i_k}^k, R_{i_k}^k)} F(x_{i_k}, \mathbf{x}_{-i_k}^k) + \frac{\gamma_{i_k}^k}{2} \|x_{i_k} - x_{i_k}^k\|^2$$

   Always have subsequence convergence

3. Blockwise $\mathbf{R}$-local prox-linear minimization:

$$x_{i_k}^{k+1} \in \operatorname*{arg\,min}_{x_i \in B(x_{i_k}^k, R_{i_k}^k)} \langle \nabla F_1(x_{i_k}^k, \mathbf{x}_{-i_k}^k), x_{i_k} - x_{i_k}^k \rangle + \frac{\gamma_{i_k}^k}{2} \|x_{i_k} - x_{i_k}^k\|^2 + F_2(x_{i_k}, \mathbf{x}_{-i_k}^k)$$

   Choose $\gamma_{i_k}^k$ to lead to a sufficient descent condition

Application

- SCAD penalty; $x \in \mathbb{R}, \gamma > 2, \lambda > 0$

$$p_{\lambda,\gamma}(x) = \begin{cases} \lambda|x| & \text{if } |x| \leq \lambda, \\ \frac{2\gamma\lambda|x| - x^2 - \lambda^2}{2(\gamma-1)} & \text{if } \lambda < |x| < \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2} & \text{if } |x| \geq \gamma\lambda \end{cases}$$

- Problem

$$\min_{\beta} Q_{\lambda,\gamma}(\beta) = \frac{1}{2}\|y - X\beta\|^2 + \sum_i p_{\lambda,\gamma}(\beta_i).$$

Compared to existing nonconvex results

- For a few applications, any local = global was recently discovered[2][3]
  Our results are weaker yet more general
- Some algorithms search all the time
  Our results only search when necessary
- Some recent results are probabilistic[4][5]
  Our results are deterministic (easier to apply)

---

[2] Ge, Huang, Jin, and Yuan [2015]
[3] Ge, Lee, and Ma [2016]
[4] Jin, Ge, Netrapalli, Kakade, and Jordan [2017]
[5] Ge, Huang, Jin, and Yuan [2015]

Example 1 Avoiding the saddle point

- Find a solution $\bar{\mathbf{x}}$ satisfying[6]

$$\|\nabla F(\mathbf{x})\| \leq \epsilon \quad \text{and} \quad \lambda_{\min}(\nabla^2 F(\mathbf{x})) \geq -\sqrt{\rho\epsilon}$$

  where ρ is the Lipschitz constant of $\nabla^2 F(\mathbf{x})$
- Problems like tensor decomposition and matrix completion enjoy strict saddle property and all local minimum is global minimum[7][8].
- Adding isotropic noise is able to find negative curvature direction with high probability[9][10].
- Probabilistic and local; mainly theoretical use

[6] Nesterov and Polyak [2006]
[7] Ge, Huang, Jin, and Yuan [2015]
[8] Ge, Lee, and Ma [2016]
[9] Jin, Ge, Netrapalli, Kakade, and Jordan [2017]
[10] Ge, Huang, Jin, and Yuan [2015]

Example 2 Flat minima in deep neural network

- Fact : flat minima are likely to have low generalization error
- Algorithm : SGD are more likely to stop in a wide valley rather than a sharp valley
- Recent Entropy-SGD[11] is a PDE based smoothing technique[12], which can make the smoothed landscape favor a flatter minima
- Their criteria of a flat minima is the behavior of eigenvalues of Hessian, which is local
- A better non-local quantity is needed to go further[13].
  Our R-local minimizer is an attempt to explore non-local property

[11] Chaudhari, Choromanska, Soatto, and LeCun [2016]
[12] Chaudhari, Oberman, Osher, Soatto, and Carlier [2017]
[13] Wu, Zhu, et al. [2017]

High-level features of our methods

- Many existing algorithms empirically work well; we add guarantees
- Finite iteration steps guarantee (deterministic)
- All sampling based method can be used in the inspection step
- We only search when and where needed
- An attempt to explore non-local properties

References:

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle pointsâĂŤonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, and Yann LeCun. Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.

Pratik Chaudhari, Adam Oberman, Stanley Osher, Stefano Soatto, and Guillame Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *arXiv preprint arXiv:1704.04932*, 2017.

Lei Wu, Zhanxing Zhu, et al. Towards understanding generalization of deep learning: Perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.