

First-Order Optimization Methods in Machine Learning

Zhouchen Lin (林宙辰)
Peking University
Aug. 27, 2016

Nonlinear Optimization: $\min_x f(x), \text{ s.t. } x \in \mathcal{C}.$

- Past (-1990s)
- Present (1990s-now)
- Future (now-)

Past (-1990s)

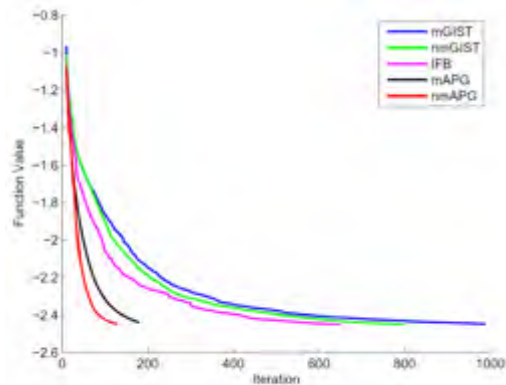
- Major theories and techniques of optimization completed
 - – 1960s
 - 1960s – 1990s, boom due to the invention of computers
- Zeroth order methods
 - Interpolation Methods
 - Pattern Search Methods
 -
- First order methods
 - Coordinate Descent
 - Conjugate Gradient
 - (Stochastic) Gradient/Subgradient Descent
 - Ellipsoid Method
 - Quasi-Newton Methods
 - (Augmented) Lagrangian Method of Multipliers
 -

Past (-1990s)

- Second order methods
 - Newton's Methods
 - Sequential Quadratic Programming
 - Interior Point Methods
 -

Why first order methods?

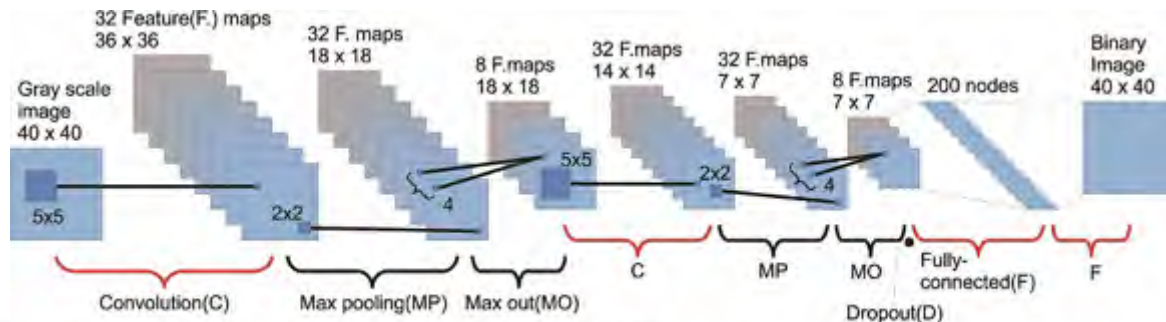
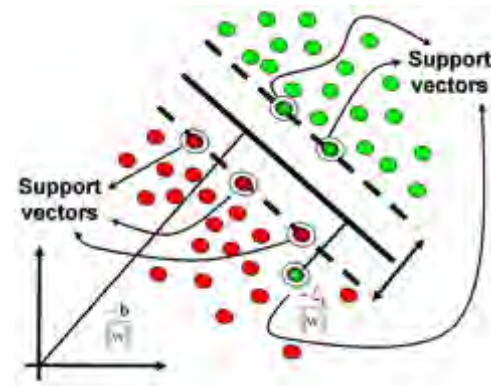
- Converge relatively fast (#iterations)
 - Acceptable accuracy for machine learning
- Relatively cheap in storage and computation (complexity in each iteration)
 - Important for big data era



(a) Objective function value v.s. iteration

Present (1990s-now)

- Revive and refine of existing techniques
 - Spiral Ascent
- Application driven
 - Support Vector Machines
 - Deep Learning
 - Big Data

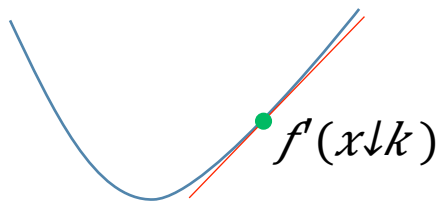


Advances in first-order methods

- Smooth \rightarrow Nonsmooth
- Convex \rightarrow Nonconvex
- Deterministic \rightarrow Stochastic
- One/Two Blocks \rightarrow Multiple Blocks
- Synchronous \rightarrow Asynchronous
- Convergence & Convergence Rate

Smooth -> Nonsmooth

- Smooth
 - Gradient

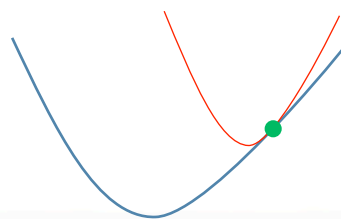


- Nonsmooth
 - Subgradient
 - Proximal Operator

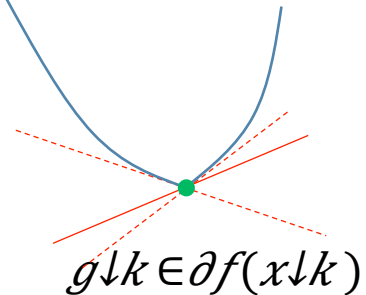
$$\min_x f(x) + \varepsilon/2 \|x - y\|_2^2$$

- Linearization

$$g(x) \leq g(x_{\downarrow k}) + \langle g'(x_{\downarrow k}), x - x_{\downarrow k} \rangle + L/2 \|x - x_{\downarrow k}\|_2^2$$



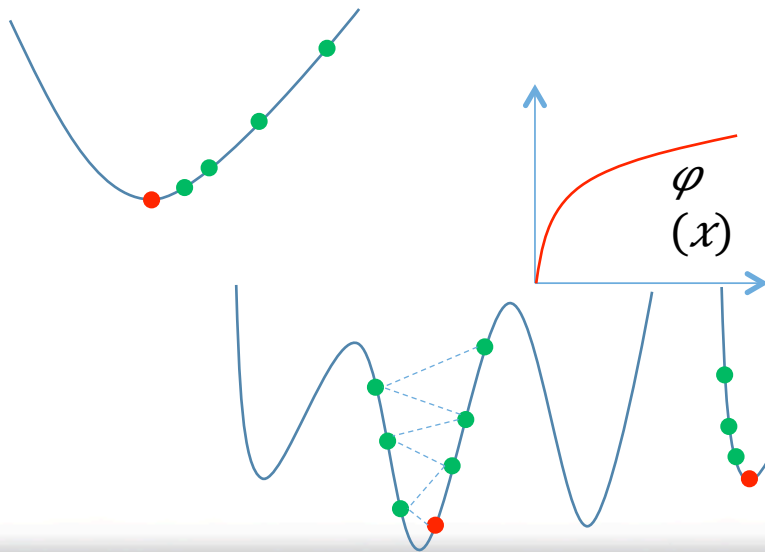
Sparsity & Low-Rankness



Convex -> Nonconvex

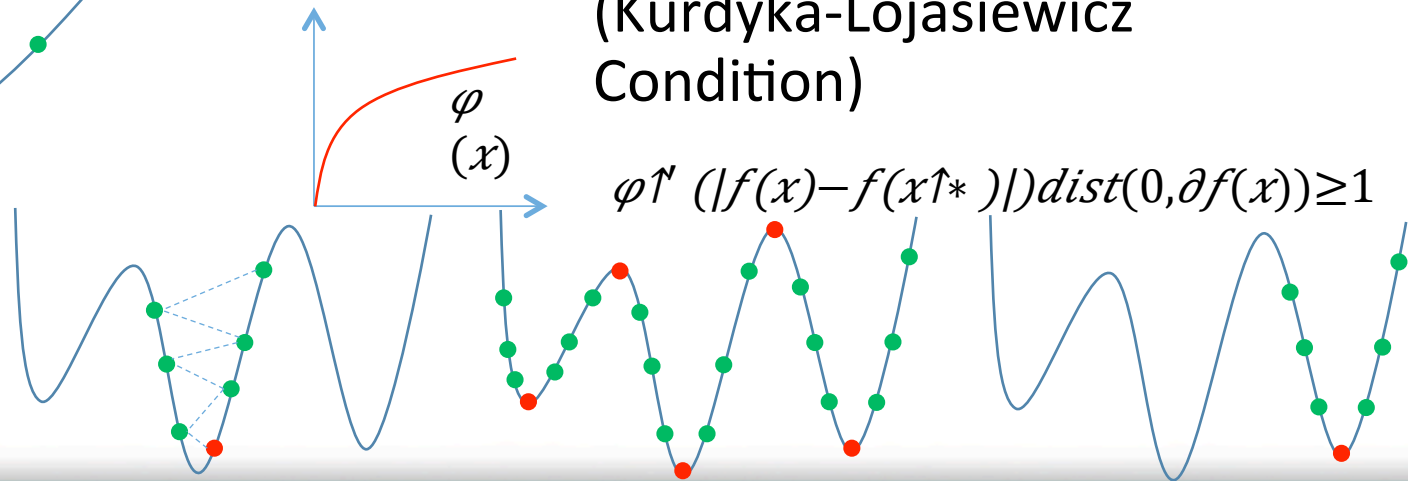
- Convex

- Global Optimal



- Nonconvex

- Nonincreasing Objective
- Cluster Points are KKT Points
- Converges to KKT Point (Kurdyka-Lojasiewicz Condition)



$$\varphi'(|f(x) - f(x^*)|) \text{dist}(0, \partial f(x)) \geq 1$$

Deterministic -> Stochastic

- Deterministic

$$f(x \downarrow k), f'(x \downarrow k)$$

- Stochastic

- Stochastic Gradient/ADMM

$$\frac{1}{n} \sum_{i=1}^n f_{i \downarrow k}'(x) \rightarrow f_{i \downarrow k}'(x)$$

- Variance Reduction
- Stochastic Matrix Computation
 - Randomized SVD

One/Two Blocks -> Multiple Blocks (ADMM)

- One/Two Blocks
 - Serial Update

$$\begin{aligned} \min & f_1(x_1) + f_2(x_2), \\ \text{s.t. } & A_1(x_1) + A_2(x_2) = z. \end{aligned}$$

- Multiple Blocks
 - Parallel Update

$$\begin{aligned} \min & f_1(x_1) + \dots + f_n(x_n), \\ \text{s.t. } & A_1(x_1) + \dots + A_n(x_n) = z. \end{aligned} \quad (n > 2)$$

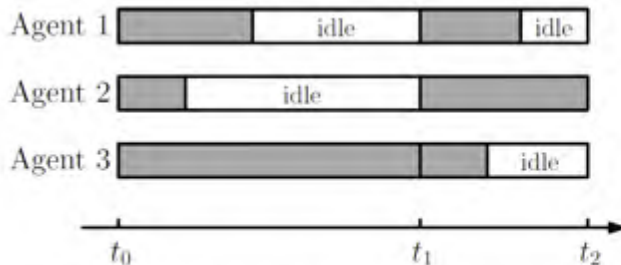
Synchronous -> Asynchronous

- Synchronous

$x_{\downarrow 1}^{\uparrow k}, x_{\downarrow 2}^{\uparrow k}, \dots, x_{\downarrow n}^{\uparrow k}$



$x_{\downarrow 1}^{\uparrow k+1}, x_{\downarrow 2}^{\uparrow k+1}, \dots,$
 $x_{\downarrow n}^{\uparrow k+1}$



(a) Sync-parallel computing

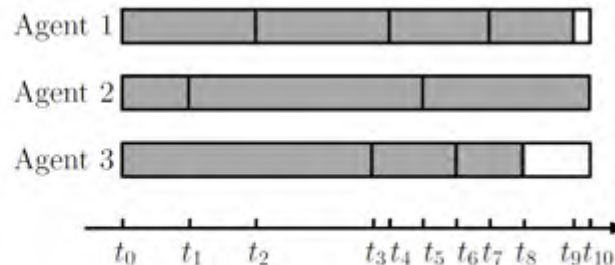
- Asynchronous

$x_{\downarrow 1}^{\uparrow k_{\downarrow 1}}, x_{\downarrow 2}^{\uparrow k_{\downarrow 2}}, \dots,$
 $x_{\downarrow n}^{\uparrow k_{\downarrow n}}$



$x_{\downarrow 1}^{\uparrow k_{\downarrow 1} + 1}, x_{\downarrow 2}^{\uparrow k_{\downarrow 2} + 1}, \dots,$
 $x_{\downarrow n}^{\uparrow k_{\downarrow n} + 1}$

Superscripts:
Iteration numbers



(b) Async-parallel computing

Convergence & Convergence Rate

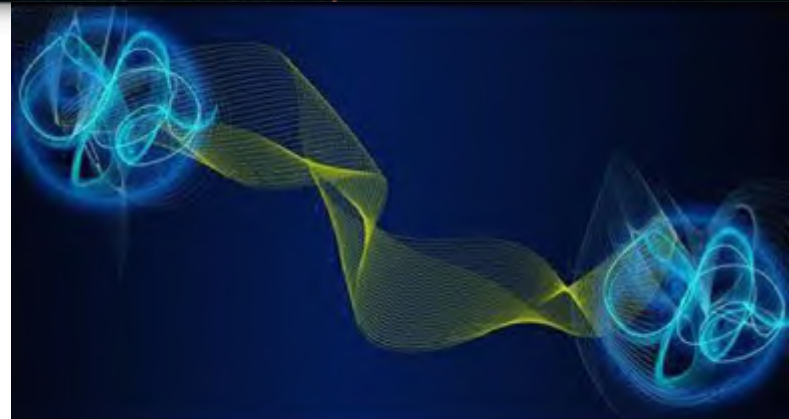
- Improved Convergence for Nonconvex Optimization (see before)
- Weaker Convergence Conditions for Convex Optimization
- Accelerated Convergence
 - Nesterov Extrapolation
 - Variance Reduction

$O(k^{\uparrow-1}) \rightarrow O(k^{\uparrow-2}), o(k^{\uparrow-1/2}) \rightarrow O(k^{\uparrow-1})$

Problem	Algorithm	Runtime
SVM	SGD	$\frac{1}{\lambda \epsilon}$
	AGD (Nesterov)	$n \sqrt{\frac{1}{\lambda \epsilon}}$
	Acc-Prox-SDCA	$(n + \min\{\frac{1}{\lambda \epsilon}, \sqrt{\frac{n}{\lambda \epsilon}}\})$
Lasso	SGD and variants	$\frac{d}{\epsilon^2} \frac{n}{\epsilon}$
	Stochastic Coordinate Descent	$n \sqrt{\frac{1}{\epsilon}}$
	FISTA	$(n + \min\{\frac{1}{\epsilon}, \sqrt{\frac{n}{\epsilon}}\})$
Ridge Regression	Acc-Prox-SDCA	$(n + \frac{1}{\lambda})$
	SGD, SDCA	$n \sqrt{\frac{1}{\lambda}}$
	AGD	$(n + \min\{\frac{1}{\lambda}, \sqrt{\frac{n}{\lambda}}\})$

Future (now-)

- Super-Large Scale Optimization
 - Fully Randomized Algorithms
 - $O(n * \text{polylog}(n))$
- Quantum Optimization
 - Quantumization of Classic Algorithms
 - Purely Quantum Algorithms



Thanks!



CCAI 2016
中国人工智能大会