# First-Order Optimization Methods in Machine Learning

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

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

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

# Past (-1990s)

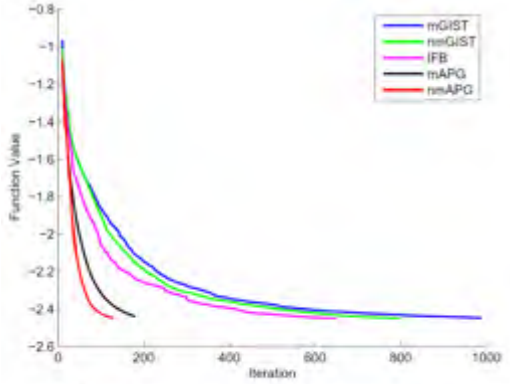
- Major theories and techniques of optimization completed
  - – 1960s
  - 1960s – 1990s, boom due to the invention of computers
- Zeroth order methods
  - Interpolation Methods
  - Pattern Search Methods
  - ……
- First order methods
  - Coordinate Descent
  - Conjugate Gradient
  - (Stochastic) Gradient/Subgradient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

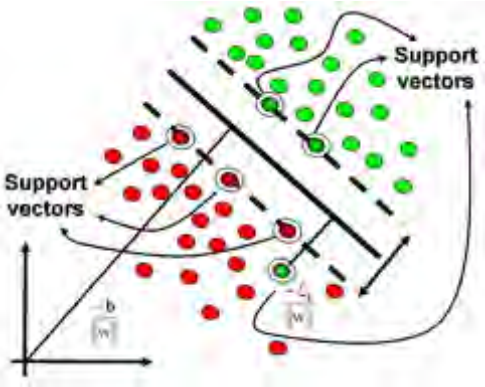
- Converge relatively fast (#iterations)
  - Acceptable accuracy for machine learning
- Relatively cheap in storage and computation (complexity in each iteration)
  - Important for big data era

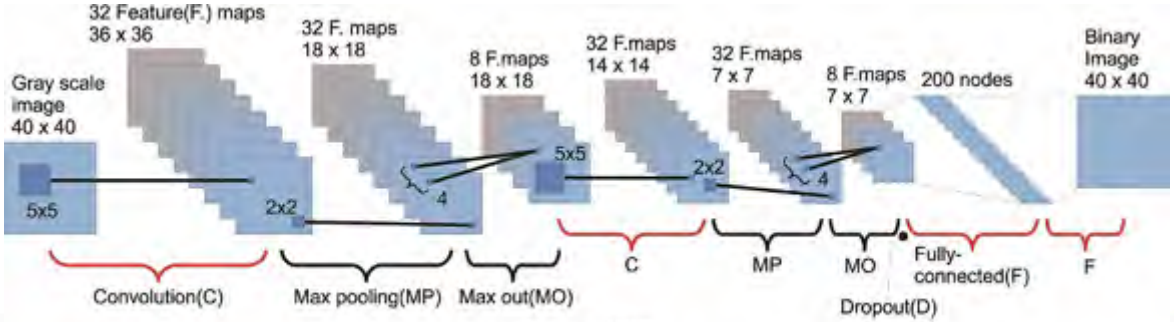



(a) Objective function value v.s. iteration

# Present (1990s-now)

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



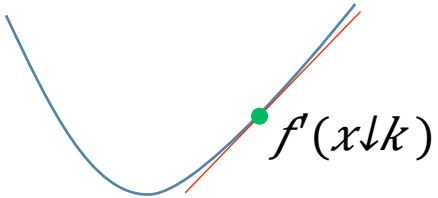



# Advances in first-order methods

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

# Smooth -> Nonsmooth

- Smooth
  - Gradient

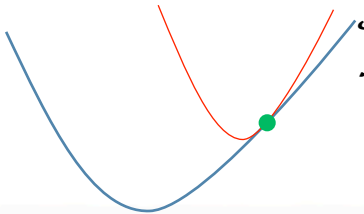



- Nonsmooth
  - Subgradient
  - Proximal Operator

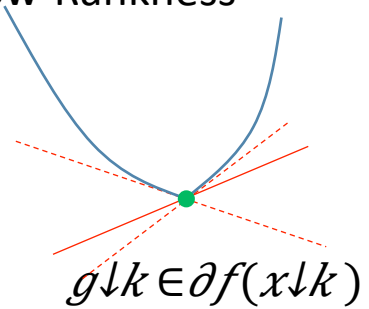
$$\min\downarrow x\, f(x)+\varepsilon/2\, \|x-y\|\uparrow 2$$

  - Linearization

$$g(x)\leq g(x\downarrow k)+\langle g'(x\downarrow k),x-x\downarrow k\rangle+L/2\, \|x-x\downarrow k\|\uparrow 2$$

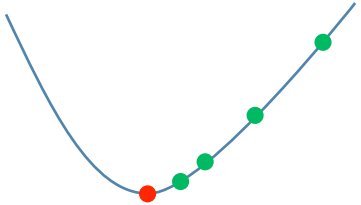


Sparsity & Low-Rankness

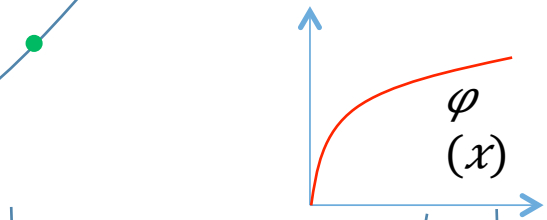


Convex -> Nonconvex

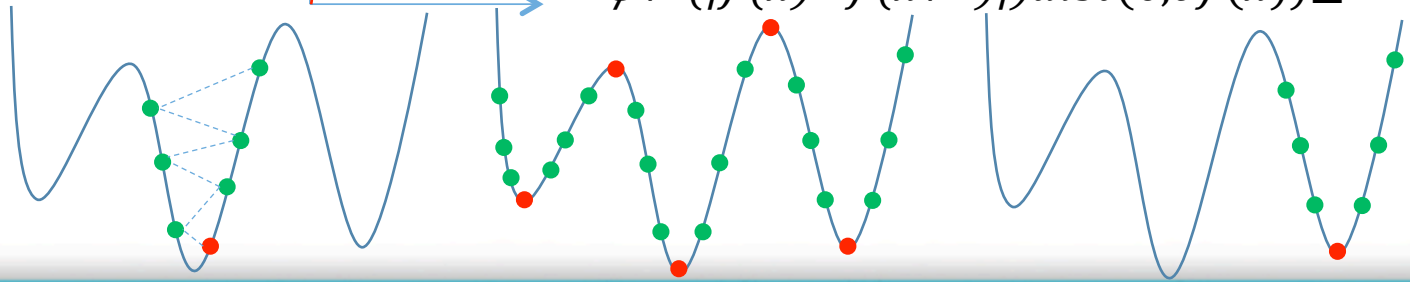
- Convex
  - Global Optimal



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




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



# Deterministic -> Stochastic

- Deterministic

  $f(x_k), f'(x_k)$

- Stochastic
  - Stochastic Gradient/ADMM

    $1/n \sum_{i=1}^{n} f_i'(x) \to f_{i_k}'(x)$

    - Variance Reduction
  - Stochastic Matrix Computation
    - Randomized SVD

# One/Two Blocks -> Multiple Blocks (ADMM)

- One/Two Blocks
  - Serial Update

$$\min\downarrow x\, f\downarrow 1\,(x\downarrow 1\,)+f\downarrow 2\,(x\downarrow 2\,),\quad s.t.\;\; A\downarrow 1\,(x\downarrow 1\,)+A\downarrow 2\,(\,x\downarrow 2\,)=z.$$

- Multiple Blocks
  - Parallel Update

$$\min\downarrow x\, f\downarrow 1\,(x\downarrow 1\,)+\cdots+f\downarrow n\,(x\downarrow n\,),\quad s.t.\;\; A\downarrow 1\,(x\downarrow 1\,)+\cdots+A\downarrow n\,(x\downarrow n\,)=z. \qquad (n>2)$$

# Synchronous -> Asynchronous

- Synchronous

$x\downarrow 1\uparrow k, x\downarrow 2\uparrow k, \cdots, x\downarrow n\uparrow k$

⇩

$x\downarrow 1\uparrow k+1, x\downarrow 2\uparrow k+1, \cdots, x\downarrow n\uparrow k+1$

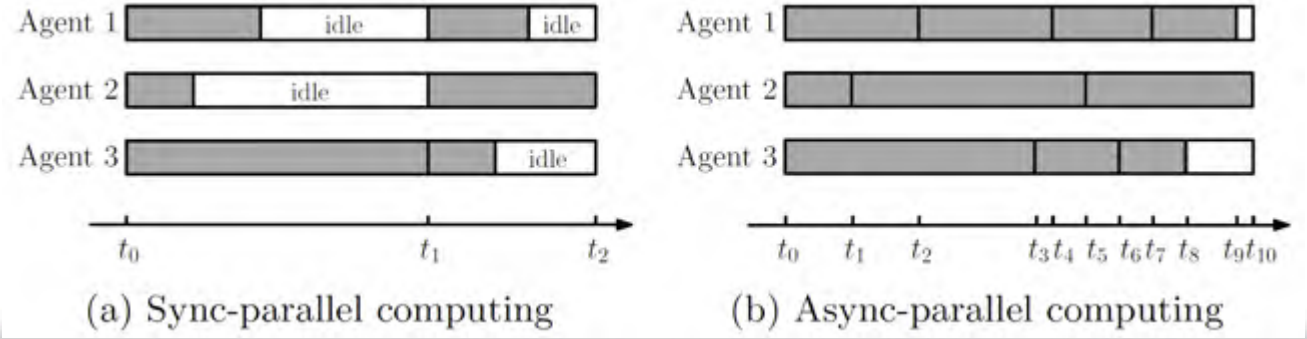
- Asynchronous

$x\downarrow 1\uparrow k\downarrow 1, x\downarrow 2\uparrow k\downarrow 2, \cdots, x\downarrow n\uparrow k\downarrow n$

⇩

$x\downarrow 1\uparrow k\downarrow 1 +1, x\downarrow 2\uparrow k\downarrow 2 +1 \cdots. x\downarrow n\uparrow k\downarrow n +1$

Superscripts:
Iteration numbers



# Convergence & Convergence Rate

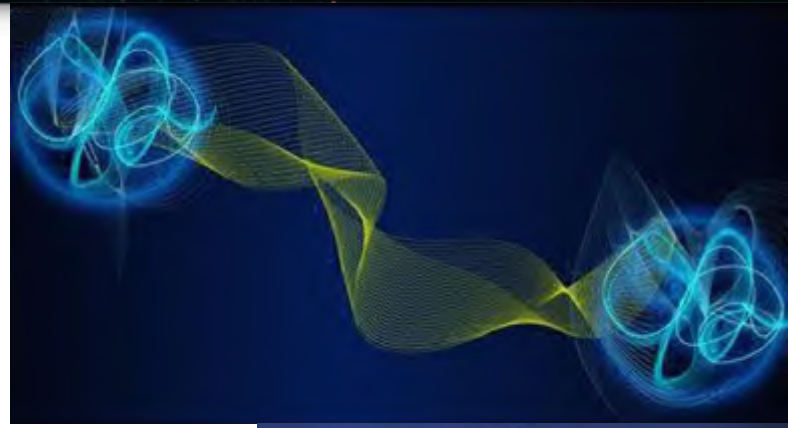
- Improved Convergence for Nonconvex Optimization (see before)
- Weaker Convergence Conditions for Convex Optimization
- Accelerated Convergence
  - Nesterov Extrapolation
  - Variance Reduction

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

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

## Future (now-)

- Super-Large Scale Optimization
  - Fully Randomized Algorithms
  - O(*n**polylog(*n*))
- Quantum Optimization
  - Quantumizaion of Classic Algorithms
  - Purely Quantum Algorithms




Thanks!
CCAI 2016
中国人工智能大会