IE598 Big Data Optimization

Summary and Outlook

Instructor: Niao He

Nov 15, 2016

Big Data, Big Picture

This Course

Big Data **Optimization**

- **Explore** modern optimization theories, algorithms, and big data applications
- **Emphasize** a deep understanding of structure of optimization problems and computation complexity of numerical algorithms
- **Expose to** the frontier of research in large-scale optimization and machine learning

Central Topics

$$
\begin{aligned}
\min_{x} \quad & f_0(x) \\
\text{s.t.} \quad & f_i(x) \leq 0, i = 1, \ldots, k \\
& h_j(x) = 0, j = 1, \ldots, \ell \\
& x \in X
\end{aligned}
$$

Large-Scale Convex Optimization

$\Uparrow$

Scalable First-Order Methods

What Did We Cover

“ The great watershed in optimization isn’t between linearity and nonlinearity, but convexity and nonconvexity.”

— R. Rockafellar, SIAM Review 1993

Fundamentals I

- Basics of Convex Optimization
  - Convex sets and convex functions
  - Operations that preserve convexity
  - Several characterizations of convex functions
  - Subgradient and subdifferential sets
  - First order optimality conditions
    - differentiable and non-differentiable cases
    - unconstrained and constrained cases
  - Lagrangian duality, saddle point, and KKT conditions
  - Convex conjugate, Fenchel duality

Fundamentals II

- Conic Optimization
  - Linear Programing (LP)
  - Second-Order Cone Programming (SOCP)
  - Semi-definite Programming (SDP)
  - Conic representable functions and sets
  - Conic duality (weak and strong duality)

- Polynomial-time Solvability
  - Interior Point Method (barrier functions, path-following)

To learn more on the theory side of convex optimization, IE 521: Convex Optimization, Spring 2017.

Highlights

- **(Convex function)** $f: R^n \rightarrow R \cup \{+\infty\}$ is convex on if $\forall x, y, \forall \lambda \in [0,1], f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$

- **(Subgradient)** g is a subgradient of a convex function f at x if
$$f(y) \geq f(x) + g^T(y - x), \forall y$$

- **(Convex program)** A local minimum is a global minimum.

- (**Optimality condition)** If f is convex and differentiable on a convex set X, then
$$x_* = argmin_{x \in X} f(x) \Leftrightarrow (x - x_*)^T \nabla f(x_*) \geq 0, \forall x \in X$$

Algorithms and Complexity I

- Smooth Convex Optimization

$$\min_{x \in X} \; f(x)$$

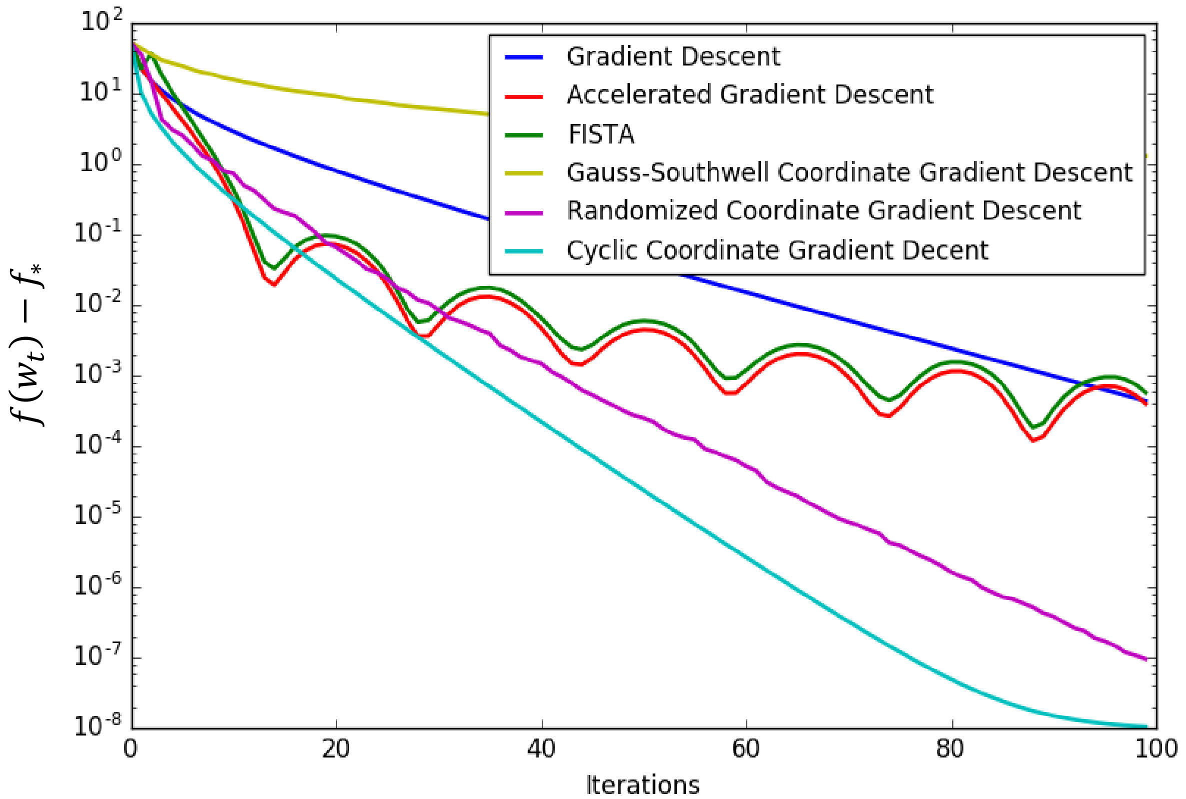
Algorithm	Iteration Complexity		Iteration Cost
	Convex	Strongly Convex	
GD	$\mathcal{O}(\frac{LD^2}{\epsilon})$	$\mathcal{O}\left(\frac{L}{\mu}\log(\frac{1}{\epsilon})\right)$	one gradient
AGD	$\mathcal{O}(\frac{\sqrt{L}D}{\sqrt{\epsilon}})$	$\mathcal{O}(\sqrt{\frac{L}{\mu}}\log(\frac{1}{\epsilon}))$	one gradient
PGD	$\mathcal{O}(\frac{LD^2}{\epsilon})$	$\mathcal{O}\left(\frac{L}{\mu}\log(\frac{1}{\epsilon})\right)$	one gradient + one projection
FW	$\mathcal{O}(\frac{LD^2}{\epsilon})$	$\mathcal{O}(\frac{LD^2}{\epsilon})$	one gradient + one linear minimization
BCGD	$\mathcal{O}(\frac{bLD^2}{\epsilon})$	$\mathcal{O}(\frac{bL}{\mu}\log(\frac{1}{\epsilon}))$	(randomized) : $O(1)$-block gradient (cyclic): $O(b)$-block gradient (Gauss Southwell): $O(b)$-block gradient

- $\boldsymbol{L}$: Lipschitz constant of $\nabla \mathrm{f}(\mathrm{x})$, $\boldsymbol{\mu}$: strongly convexity, $\boldsymbol{D}$: either $||x_0 - x_*||_2$ or diameter of set X

Example

- Logistic Regression

$$\min_w f(w) := \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i w^T x_i)) + \frac{\lambda}{2}\|w\|_2^2$$



Algorithms and Complexity II

- Nonsmooth Convex Optimization

$$\min_{x \in X} \; f(x)$$

Algorithm	Iteration Complexity (Convex Case)	Iteration Cost
Subgradient Descent	$\mathcal{O}\left(\frac{M^2_{\|\cdot\|_2}(f)\cdot \max_{x,y\in X}\|x-y\|_2^2}{\epsilon^2}\right)$	one subgradient one projection
Mirror Descent	$\mathcal{O}\left(\frac{M^2_{\|\cdot\|_*}(f)\cdot \max_{x,y\in X} V(x,y)}{\epsilon^2}\right)$	one subgradient one prox-mapping
Proximal Point Algorithm	$\mathcal{O}\left(\frac{\|x_0-x_*\|_2^2}{\epsilon}\right)$	one proximal operator
Acc Proximal Point Algorithm	$\mathcal{O}\left(\frac{\|x_0-x_*\|_2}{\sqrt{\epsilon}}\right)$	one proximal operator

- $\boldsymbol{V(x,y)}$: Bregman distance w.r.t. some norm $\|\cdot\|$ defined on X, $\boldsymbol{M}$: Lipschitz constant of $f(x)$

Algorithms and Complexity III

- Nonsmooth Convex Optimization

$$\min_{x \in X} f(x) := \max_{y \in Y}\{\langle Ax + b, y\rangle - \phi(y)\}$$

Algorithm	Iteration Complexity (Convex Case)	Iteration Cost
Nesterov's Smoothing + GD	$\mathcal{O}\left(\frac{\|A\|^2 D_X^2 D_Y^2}{\epsilon^2}\right)$	one gradient of smoothed objective
Nesterov's Smoothing + AGD	$\mathcal{O}\left(\frac{\|A\| D_X D_Y}{\epsilon}\right)$	one gradient of smoothed objective
Mirror Prox	$\mathcal{O}\left(\frac{L \cdot \max_{z,z' \in X \times Y} V(z,z')}{\epsilon}\right)$	two gradients and two prox-mappings

- $\boldsymbol{V(z,z')}$: Bregman distance w.r.t. some norm $\|\cdot\|$ defined on $X \times Y$, $\boldsymbol{D_X}, \boldsymbol{D_Y}$: diameter of sets X and Y
- $\boldsymbol{L}$: Lipschitz constant of the gradient of the saddle function

Algorithms and Complexity IV

- Nonsmooth Convex Optimization

$$\min_{x} \; f(x) + g(x)$$

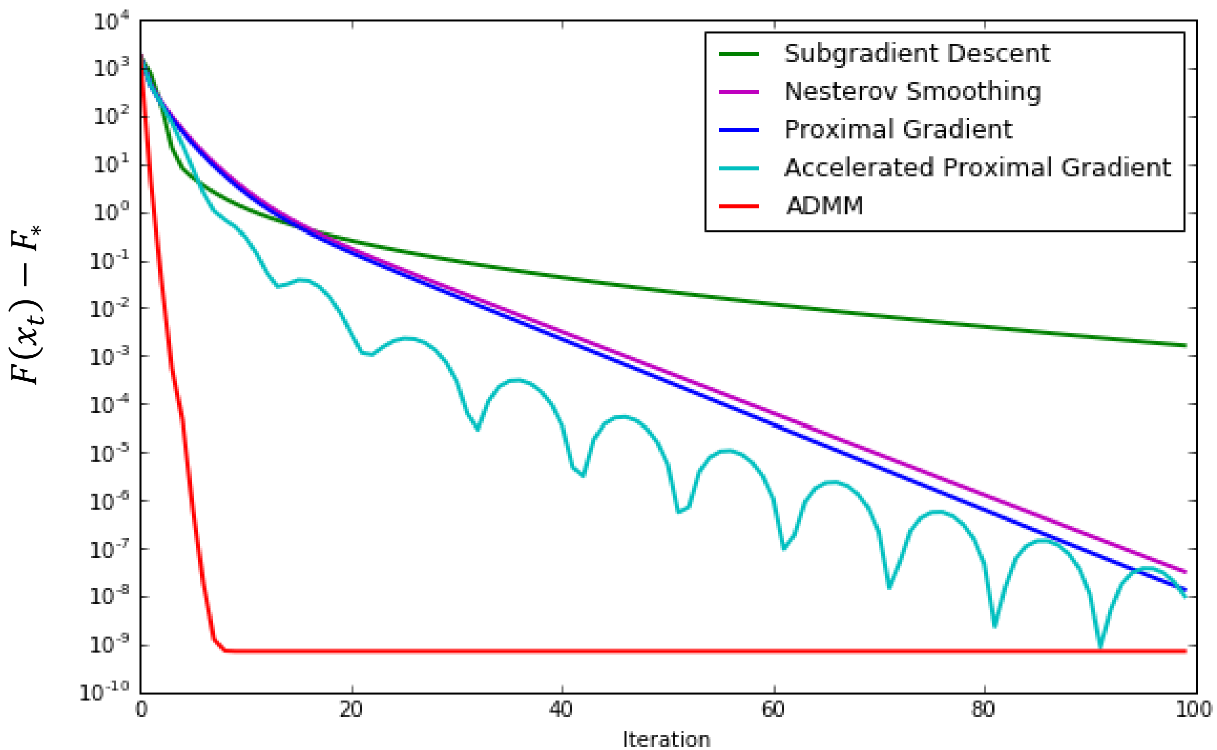
Algorithm	Iteration Complexity (Convex Case)	Iteration Cost
Proximal Gradient	$\mathcal{O}\left(\frac{L\|x_0 - x_*\|_2^2}{\epsilon}\right)$	one gradient of f one proximal operator of g
Accelerated Proximal Gradient	$\mathcal{O}\left(\frac{\sqrt{L}\|x_0 - x_*\|_2}{\sqrt{\epsilon}}\right)$	one gradient of f one proximal operator of g
Douglas-Rachford Splitting (special case of ADMM)	$\mathcal{O}\left(\frac{\|x_0 - x_*\|_2^2}{\epsilon}\right)$	one proximal operator of f one proximal operator of g
Krasnosel'skii-Mann(KM) (generalization)	$\mathcal{O}\left(\frac{\|x_0 - x_*\|_2^2}{\epsilon}\right)$	one proximal operator of f one proximal operator of g

- $\boldsymbol{L}$: Lipschitz constant of $\nabla f(x)$

Example

- LASSO

$$\min_x \frac{1}{2}||Ax - b||_2^2 + \lambda ||x||_1$$



Algorithms and Complexity V

- Stochastic Convex Optimization

$$\min_{x \in X} \; f(x) = \mathbb{E}[F(x,\xi)]$$

Approach	Sample Complexity
Sample Average Approximation (SAA)	$\mathcal{O}\left(\frac{\max_{x\in X} Var[F(x,\xi)]}{\epsilon^2}\right)$
Stochastic Approximation (SA) (when f is μ-strongly convex)	$\mathcal{O}\left(\frac{\max_{x\in X} \mathbb{E}[\lVert F'(x,\xi)\rVert_2^2]}{\mu^2\epsilon}\right)$
Mirror Descent SA (when f is general convex)	$\mathcal{O}\left(\frac{\max_{x\in X} \mathbb{E}[\lVert F'(x,\xi)\rVert_*^2]\cdot\max_{x,y\in X} V(x,y)}{\epsilon^2}\right)$

  - $\boldsymbol{V(x,y)}$: Bregman distance w.r.t. some norm $||\cdot||$ defined on X

Algorithms and Complexity VI

- Finite Sum of Convex Functions

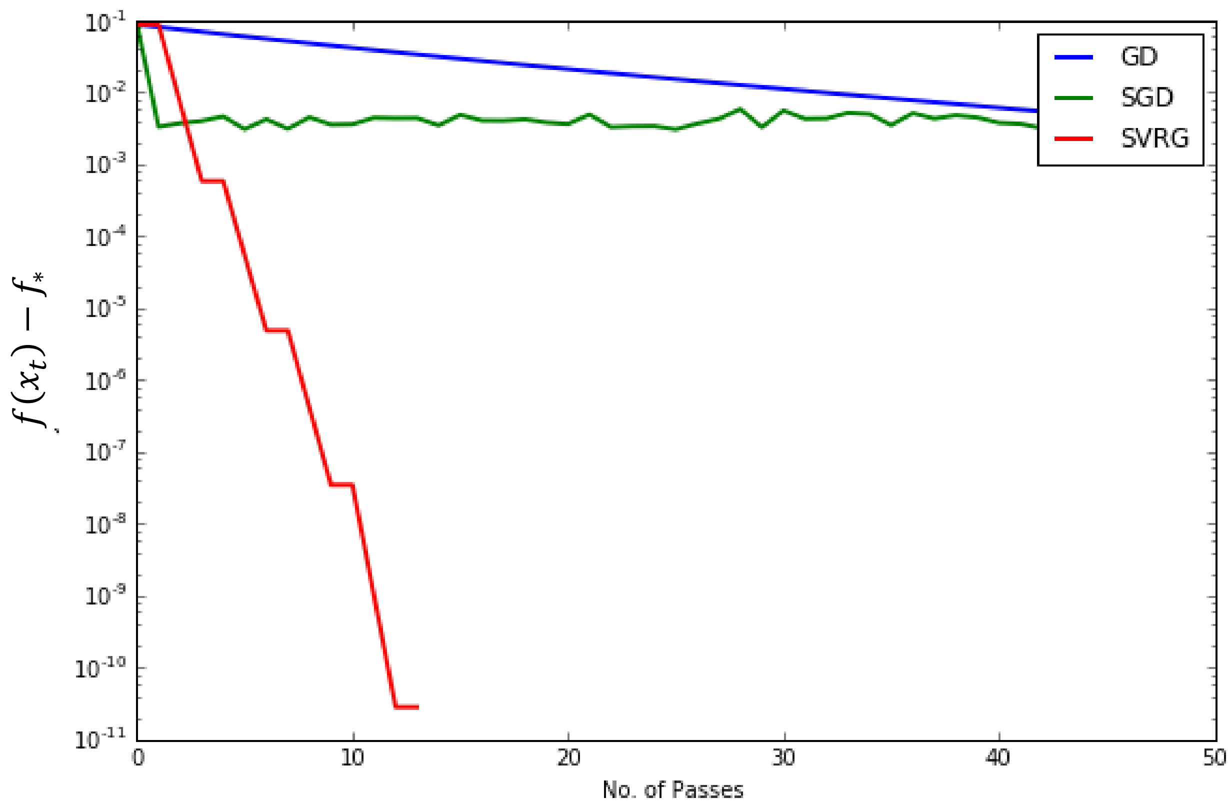
$$\min_x \; f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Algorithm	Iteration Complexity (Smooth + Strongly Convex)	Iteration Cost
GD	$\mathcal{O}\left(\frac{L}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$	$O(n)$ gradient
SGD	$\mathcal{O}\left(\frac{L}{\mu^2\epsilon}\right)$	$O(1)$ gradient
SVRG/S2GD	$\mathcal{O}\left(\log\left(\frac{1}{\epsilon}\right)\right)$	$O(n+\frac{L}{\mu})$ gradient
SAG/SAGA	$\mathcal{O}\left(\max(n,\frac{L}{\mu})\log\left(\frac{1}{\epsilon}\right)\right)$	$O(1)$ gradient $O(n)$ memory

Example

- Large-scale Logistic Regression

$$\min_{w} f(w) := \frac{1}{n}\sum_{i=1}^{n} \log(1+\exp(-y_i w^T x_i)) + \frac{\lambda}{2}\|w\|_2^2$$



What We Did Not Cover

“Optimization hinders evolution.”

-- Alan J. Perlis, 1982

More First-Order Algorithms

- **Nonsmooth Optimization Algorithms**
  - Bundle methods (e.g. the level method)
  - Primal-dual methods
  - Composite Mirror Descent/Mirror Prox
  - …

- **Stochastic Optimization Algorithms**
  - Dual averaging method
  - Stochastic Frank Wolfe algorithms
  - Stochastic dual coordinate ascent
  - Stochastic ADMM algorithm
  - …

Beyond First-Order Algorithms

- **Second-Order Methods**
  - Newton method
  - (stochastic) Quasi-Newton methods
  - Gauss-Newton method
  - Natural Gradient method
  - …
- **Zero-Order Methods (Derivative-free)**
  - Fast Differentiation technique
  - Gaussian smoothing
  - Random search
  - …

Beyond Black-Box

Methods with linear dimension-dependent convergence

- Cutting plane methods
- Center-of-Gravity method
- Inner and Outer Ellipsoid method
- Interior Point Method

Parallel and Distributed Algorithms

Many of the algorithms we learnt can be modified to take advantage of parallel processors and distributed machines

- Distributed ADMM
- Async-ADMM
- Hogwild!
- Downpour SGD
- Distributed dual averaging
- Gossip algorithms

More on Convex Optimization

- Problems with Convex Structure
  - Convex Minimization
  - Convex-Concave Saddle Point Problems
  - Variational Inequalities
  - Convex Nash Equilibrium
  - Monotone Inclusion Problems

- Convex Optimization under Hilbert spaces

- Online Convex Optimization

More Applications in ML

Aside from supervised learning, many other tasks in ML are also convex problems

- Boosting
- Bayesian Inference
- Reinforcement Learning
- Recommendation Systems
- Social Network Estimation
- …

Beyond Convex Optimization

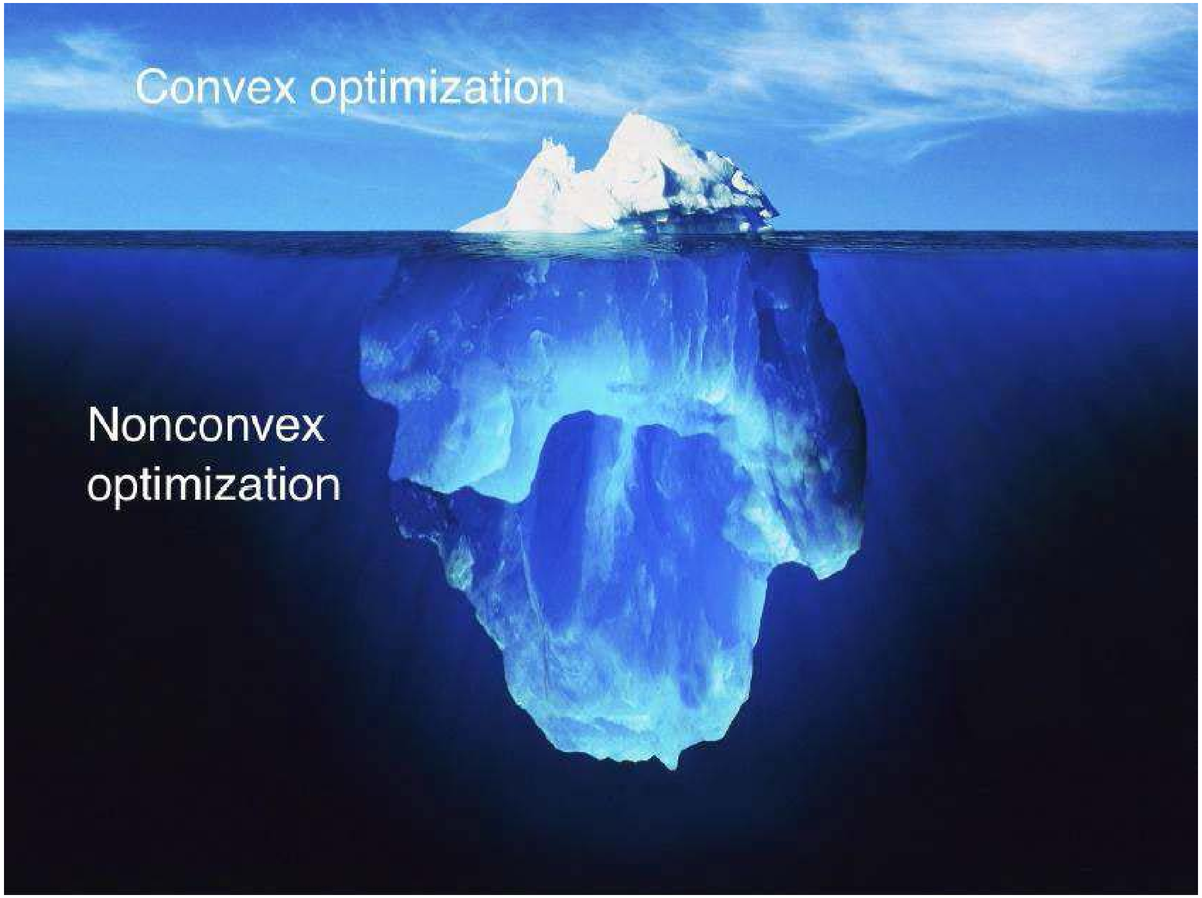


Image source: https://www.facebook.com/nonconvex

Non-Convex Problems

Many practical problems are non-convex and are hard to solve.

- **Nonlinear Optimization**
  - Polynomial optimization
  - Convex equality constraints
  - Eigenvalue problems
- **Integer and Combinatorial Optimization**
- **Optimization under Uncertainty**
  - Robust Optimization
  - Chance Constrained Programming
  - Multi-Stage Stochastic Programming

Non-Convex Applications in ML

Lots of problems in machine learning are indeed non-convex, for instance

- Deep Learning
- Clustering (K-means, PCA, etc)
- Graphical Models (MRF, HMM, etc)
- Multi-class Classification
- Sparsity learning with non-convex regularization

Non-Convex Algorithms

- Converging to stationary points
  - Many algorithms from the convex world still apply but with weaker convergence
  - For example, GD, FW, SGD, SVRG, etc.
- Escaping from saddle points
  - Restarting
  - Using noisy gradient
  - Using Hessian information
- Converging to global optimum
  - Proven to be possible for several family of problems

Read more : http://www.offconvex.org/