IE598 Big Data Optimization

Summary and Outlook

Instructor: Niao He
Nov 15, 2016
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

\[
\begin{aligned}
\min_x & \quad f_0(x) \\
\text{s.t.} & \quad f_i(x) \leq 0, \ i = 1, \ldots, k \\
& \quad h_j(x) = 0, \ j = 1, \ldots, \ell \\
& \quad x \in X
\end{aligned}
\]
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.
• **(Convex function)** $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ 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 $x_* = \arg\min_{x \in X} f(x) \Leftrightarrow (x - x_*)^T \nabla f(x_*) \geq 0, \forall x \in X$
## Smooth Convex Optimization

\[
\min_{x \in X} f(x)
\]

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

- \( L \): Lipschitz constant of \( \nabla f(x) \)
- \( \mu \): strongly convexity
- \( D \): either \( \|x_0 - x_*\|_2 \) or diameter of set \( X \)
Example

- Logistic Regression

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

\[
\min_{x \in X} f(x)
\]

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Iteration Complexity (Convex Case)</th>
<th>Iteration Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgradient Descent</td>
<td>( \mathcal{O}\left(\frac{M^2|\cdot|<em>2(f)\cdot \max</em>{x,y \in X} |x-y|_2^2}{\epsilon^2}\right) )</td>
<td>one subgradient one projection</td>
</tr>
<tr>
<td>Mirror Descent</td>
<td>( \mathcal{O}\left(\frac{M^2|\cdot|<em>\infty(f) \cdot \max</em>{x,y \in X} V(x,y)}{\epsilon^2}\right) )</td>
<td>one subgradient one prox-mapping</td>
</tr>
<tr>
<td>Proximal Point Algorithm</td>
<td>( \mathcal{O}\left(\frac{|x_0 - x_*|_2^2}{\epsilon}\right) )</td>
<td>one proximal operator</td>
</tr>
<tr>
<td>Acc Proximal Point Algorithm</td>
<td>( \mathcal{O}\left(\frac{|x_0 - x_*|_2}{\sqrt{\epsilon}}\right) )</td>
<td>one proximal operator</td>
</tr>
</tbody>
</table>

- \( V(x,y) \): Bregman distance w.r.t. some norm \( \| \cdot \| \) defined on \( X \), \( M \): Lipschitz constant of \( f(x) \)
• Nonsmooth Convex Optimization

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Iteration Complexity</th>
<th>Iteration Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nesterov’s Smoothing + GD</td>
<td>( \mathcal{O} \left( \frac{|A|^2 D_x^2 D_y^2}{\epsilon^2} \right) )</td>
<td>one gradient of smoothed objective</td>
</tr>
<tr>
<td>Nesterov’s Smoothing + AGD</td>
<td>( \mathcal{O} \left( \frac{|A| D_x D_y}{\epsilon} \right) )</td>
<td>one gradient of smoothed objective</td>
</tr>
<tr>
<td>Mirror Prox</td>
<td>( \mathcal{O} \left( \frac{L \cdot \max_{z,z' \in X \times Y} V(z,z')}{\epsilon} \right) )</td>
<td>two gradients and two prox-mappings</td>
</tr>
</tbody>
</table>

- \( V(z,z') \): Bregman distance w.r.t. some norm \( \| \cdot \| \) defined on \( X \times Y \), \( D_x, D_y \): diameter of sets \( X \) and \( Y \)
- \( L \): Lipschitz constant of the gradient of the saddle function
### Nonsmooth Convex Optimization

\[
\min_x f(x) + g(x)
\]

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Iteration Complexity (Convex Case)</th>
<th>Iteration Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proximal Gradient</td>
<td>( \mathcal{O}\left( \frac{L|x_0-x_*|_2^2}{\epsilon} \right) )</td>
<td>one gradient of ( f ), one proximal operator of ( g )</td>
</tr>
<tr>
<td>Accelerated Proximal Gradient</td>
<td>( \mathcal{O}\left( \frac{\sqrt{L}|x_0-x_*|_2}{\sqrt{\epsilon}} \right) )</td>
<td>one gradient of ( f ), one proximal operator of ( g )</td>
</tr>
<tr>
<td>Douglas-Rachford Splitting</td>
<td>( \mathcal{O}\left( \frac{|x_0-x_*|_2^2}{\epsilon} \right) )</td>
<td>one proximal operator of ( f ), one proximal operator of ( g )</td>
</tr>
<tr>
<td>Krasnosel’ski-Mann(KM) (generalization)</td>
<td>( \mathcal{O}\left( \frac{|x_0-x_*|_2^2}{\epsilon} \right) )</td>
<td>one proximal operator of ( f ), one proximal operator of ( g )</td>
</tr>
</tbody>
</table>

- \( L \) : Lipschitz constant of \( \nabla f(x) \)
Example

• LASSO

$$\min_{x} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1$$
**Stochastic Convex Optimization**

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

<table>
<thead>
<tr>
<th>Approach</th>
<th>Sample Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Average Approximation (SAA)</td>
<td>( O \left( \frac{\max_{x \in X} \text{Var}[F(x, \xi)]}{\epsilon^2} \right) )</td>
</tr>
<tr>
<td><strong>Stochastic Approximation (SA)</strong></td>
<td>( O \left( \frac{\max_{x \in X} \mathbb{E}[|F'(x, \xi)|_2^2]}{\mu^2 \epsilon} \right) )</td>
</tr>
<tr>
<td>(when ( f ) is ( \mu )-strongly convex)</td>
<td></td>
</tr>
<tr>
<td><strong>Mirror Descent SA</strong></td>
<td>( O \left( \frac{\max_{x \in X} \mathbb{E}[|F'(x, \xi)|<em>2^2] \cdot \max</em>{x, y \in X} V(x, y)}{\epsilon^2} \right) )</td>
</tr>
<tr>
<td>(when ( f ) is general convex)</td>
<td></td>
</tr>
</tbody>
</table>

- \( V(x, y) \) : Bregman distance w.r.t. some norm \( \| \cdot \| \) defined on \( X \)
**Finite Sum of Convex Functions**

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Iteration Complexity (Smooth + Strongly Convex)</th>
<th>Iteration Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD</td>
<td>(O\left(\frac{L}{\mu} \log \left(\frac{1}{\epsilon}\right)\right))</td>
<td>(O(n)) gradient</td>
</tr>
<tr>
<td>SGD</td>
<td>(O\left(\frac{L}{\mu^2 \epsilon}\right))</td>
<td>(O(1)) gradient</td>
</tr>
<tr>
<td>SVRG/S2GD</td>
<td>(O\left(\log \left(\frac{1}{\epsilon}\right)\right))</td>
<td>(O(n + \frac{L}{\mu})) gradient</td>
</tr>
<tr>
<td>SAG/SAGA</td>
<td>(O\left(\max(n, \frac{L}{\mu}) \log \left(\frac{1}{\epsilon}\right)\right))</td>
<td>(O(1)) gradient, (O(n)) memory</td>
</tr>
</tbody>
</table>
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
- ...

24
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/](http://www.offconvex.org/)