First order methods

FOR CONVEX OPTIMIZATION

Saketh (IIT Bombay)
Topics

• Part – I
  • Optimal methods for unconstrained convex programs
    • Smooth objective
    • Non-smooth objective

• Part – II
  • Optimal methods for constrained convex programs
    • Projection based
    • Frank-Wolfe based
    • Functional constraint based
  • Prox-based methods for structured non-smooth programs
Constrained Optimization - Illustration
Constrained Optimization - Illustration

\[ x^* \text{ is optimal} \iff \nabla f(x^*)^T u \geq 0 \ \forall \ u \in T_F(x^*) \]
Two Strategies

- Stay feasible and minimize
  - Projection based
  - Frank-Wolfe based
Two Strategies

• Alternate between
  • Minimization
  • Move towards feasibility set
Projection Based Methods

CONSTRAINED CONVEX PROGRAMS
Projected Gradient Method

\[
\min_{x \in X} f(x) \quad X \text{ is closed convex}
\]

\[
x_{k+1} = \operatorname{argmin}_{x \in X} f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2
\]

\[
= \operatorname{argmin}_{x \in X} \|x - (x_k - s_k \nabla f(x_k))\|^2
\]

\[
\equiv \Pi_X(x_k - s_k \nabla f(x_k))
\]
Projected Gradient Method

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

\[ x_{k+1} = \arg\min_{x \in X} f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2 \]

\[ = \arg\min_{x \in X} \|x - (x_k - s_k \nabla f(x_k))\|^2 \]

\[ \equiv \Pi_X(x_k - s_k \nabla f(x_k)) \]
Projected Gradient Method

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

\[
x_{k+1} = \arg\min_{x \in X} f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2
\]

\[
= \arg\min_{x \in X} \|x - (x_k - s_k \nabla f(x_k))\|^2
\]

\[
\equiv \Pi_X (x_k - s_k \nabla f(x_k))
\]

X is simple: oracle for projections
Projected Gradient Method

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

\[
x_{k+1} = \arg\min_{x \in X} f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2
\]

\[
= \arg\min_{x \in X} \|x - (x_k - s_k \nabla f(x_k))\|^2
\]

\[
\equiv \Pi_X (x_k - s_k \nabla f(x_k))
\]
Will it work?

- \[ \|x_{k+1} - x^*\|^2 = \|\Pi_X(x_k - s_k \nabla f(x_k)) - x^*\|^2 \leq \| (x_k - s_k \nabla f(x_k)) - x^*\|^2 \] (Why?)

- Remaining analysis exactly same (smooth/non-smooth)

- Analysis a bit more involved for projected accelerated gradient
  - Define gradient map: \[ h(x_k) \equiv \frac{x_k - \Pi_X(x_k - s_k \nabla f(x_k))}{s_k} \]
  - Satisfies same fundamental properties as gradient!
Will it work?

\[ \|x_{k+1} - x^*\|^2 = \|\Pi_X(x_k - s_k \nabla f(x_k)) - x^*\|^2 \]
\[ \leq \| (x_k - s_k \nabla f(x_k)) - x^*\|^2 \] (Why?)

- Remaining analysis exactly same (smooth/non-smooth)

- Analysis a bit more involved for projected accelerated gradient
  - Define gradient map: \[ h(x_k) \equiv \frac{x_k - \Pi_X(x_k - s_k \nabla f(x_k))}{s_k} \]
  - Satisfies same fundamental properties as gradient!
Simple sets

• Non-negative orthant
• Ball, ellipse
• Box, simplex
• Cones
• PSD matrices
• Spectrahedron
Summary of Projection Based Methods

• Rates of convergence remain exactly same
• Projection oracle needed (simple sets)
  • Caution with non-analytic cases
Frank-Wolfe Methods

CONSTRAINED CONVEX PROGRAMS
Avoid Projections

\[ y_{k+1} = \arg\min_{x \in X} f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \| x - x_k \|^2 \]

\[ = \arg\min_{x \in X} \nabla f(x_k)^T x \]  
(Support Function)

• Restrict moving far away:
  • \[ x_{k+1} = s_k y_{k+1} + (1 - s_k) x_k \]
Avoid Projections [FW59]

\[ y_{k+1} = \arg\min_{x \in X} f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2 \]

\[ = \arg\min_{x \in X} \nabla f(x_k)^T x \]  
(Support Function)

• Restrict moving far away:
  • \( x_{k+1} \equiv s_k y_{k+1} + (1 - s_k)x_k \)
Illustration

\[ f(x) \]

\[ y \]

\[ x^+ \]

\[ -\nabla f(x) \]

[Mart Jaggi, ICML 2014]
On Conjugates and Support Functions

• Convex $f$ is point-wise maximum of affine minorants
• Provides dual definition:
  • $f(x) = \max_{y \in Y} a^T y x - b y$, equivalently:
  • $\exists f^* \ni f(x) = \max_{y \in \text{dom } f^*} y^T x - f^*(y)$
  • $f^*$ is called conjugate or Fenchel dual
• If $f^*(y)$ is indicator of set $S$ we get conic $f$:
  • $f(x) = \max_{y \in S} y^T x$
On Conjugates and Support Functions

- Convex $f$ is point-wise maximum of affine minorants
- Provides dual definition:
  - $f(x) = \max_{y \in Y} a_y^T x - b_y$, equivalently:
  - $\exists f^* \exists f(x) = \max_{y \in \text{dom} f^*} y^T x - f^*(y)$
  - $f^*$ is called conjugate or Fenchel dual or Legendre transformation ($f^{**} = f$).
- If $f^*(y)$ is indicator of set $S$ we get conic $f$:
  - $f(x) = \max_{y \in S} y^T x$
On Conjugates and Support Functions

- Convex $f$ is point-wise maximum of affine minorants
- Provides dual definition:
  - $f(x) = \max_{y \in Y} a_y^T x - b_y$, equivalently:
  - $\exists f^* \ni f(x) = \max_{y \in \text{dom } f^*} y^T x - f^*(y)$
  - $f^*$ is called conjugate or Fenchel dual or Legendre transformation ($f^{**} = f$).
- If $f^*(y)$ is indicator of set $S$ we get conic $f$:
  - $f(x) = \max_{y \in S} y^T x$
- If $S$ is a norm ball, we get dual norm
Connection with sub-gradient

Let,

- $y^* \in \arg\max_{y \in \text{dom } f} y^T x - f(y)$ i.e., $f^*(x) + f(y^*) = x^T y^*$

- Then $y^*$ must be a sub-gradient of $f^*$ at $x$
  - dual form exposes sub-gradient

- If $f^*(y)$ is indicator of set $S$ we get conic $f$:
  - $f(x) = \max_{y \in S} y^T x$
Conjugates e.g.

<table>
<thead>
<tr>
<th>$f(x)$</th>
<th>$f^*(x)$</th>
<th>Projection?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$|x|_p$</td>
<td>$|x|_{\frac{p}{p-1}}$</td>
<td>No ($p \notin {1, 2, \infty}$)</td>
</tr>
<tr>
<td>$|\sigma(X)|_p$</td>
<td>$|\sigma(X)|_{\frac{p}{p-1}}$</td>
<td>No ($p \notin {1, 2, \infty}$)</td>
</tr>
</tbody>
</table>

- $\|x\|_1$ Projection, conjugate $= O(n \log n), O(n)$
- $\|\sigma(X)\|_1$ Projection, conjugate $= \text{Full, First SVD}$
Rate of Convergence

**Theorem [Ma11]:** If $X$ is compact convex set and $f$ is smooth with const. $L$, and $s_k = \frac{2}{k+2}$, then the iterates generated by Frank-Wolfe satisfy:

$$f(x_k) - f(x^*) \leq \frac{4L d(X)^2}{k + 2}.$$  

**Proof Sketch:**

- $f(x_{k+1}) \leq f(x_k) + s_k \nabla f(x_k)^T(y_{k+1} - x_k) + \frac{s_k^2 L}{2} d(X)^2$

- $\Delta_{k+1} \leq (1 - s_k) \Delta_k + \frac{s_k^2 L}{2} d(X)^2$ (Solve recursion)
Rate of Convergence

**Theorem [Ma11]**: If $X$ is compact convex set and $f$ is smooth with const. $L$, and $s_k = \frac{2}{k+2}$, then the iterates generated by Frank-Wolfe satisfy:

$$f(x_k) - f(x^*) \leq \frac{4L \cdot d(X)^2}{k + 2}.$$

**Proof Sketch**:

- $f(x_{k+1}) \leq f(x_k) + s_k \nabla f(x_k)^T (y_{k+1} - x_k) + \frac{s_k^2 L}{2} d(X)^2$
- $\Delta_{k+1} \leq (1 - s_k) \Delta_k + \frac{s_k^2 L}{2} d(X)^2$ (Solve recursion)
Sparse Representation – Optimality

• If $x_0 = 0$ and domain is $l_1$ ball, $x_k \in R^{k,n}$
  • We get exact sparsity! (unlike proj. grad.)
• Sparse representation by extreme points

• $\epsilon \approx O\left(\frac{k \cdot d(x)^2}{L}\right)$ need atleast $k$ non-zeros [Ma11]
• Optimal in terms of accuracy-sparsity trade-off
  • Not in terms of accuracy-iterations
Sparse Representation – Optimality

- If $x_0 = 0$ and domain is $l_1$ ball, $x_k \in \mathbb{R}^{k,n}$
  - We get exact sparsity! (unlike proj. grad.)

- Sparse representation by extreme points

- $\epsilon \approx O^\left(L d(X)^2 / k \right)$ need atleast $k$ non-zeros [Ma11]

- Optimal in terms of accuracy-sparsity trade-off
  - Not in terms of accuracy-iterations
Summary comparison of always feasible methods

<table>
<thead>
<tr>
<th>Property</th>
<th>Projected Gr.</th>
<th>Frank-Wolfe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rate of convergence</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Sparse Solutions</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Iteration Complexity</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Affine Invariance</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>
Functional Constrained BASED METHODS
Assumptions

\[
\min_{x \in \mathbb{R}^n} f_0(x) \\
\text{s.t. } f_i(x) \leq 0 \quad \forall \ i = 1, \ldots, n
\]

• All \( f_0, f_i \) are smooth
• L is max. const. among all
Algorithm

At iteration $1 \leq k \leq N$:

- Check if $f_i(x_k) \leq \frac{R}{\sqrt{N}} \|\nabla f_i(x_k)\| \forall i$
  - If yes, then “productive” step: $i(k) = 0$
  - If no, then “non-productive” step: $i(k)$ set to a violator

- $x_{k+1} = x_k - \frac{R}{\sqrt{N} \|\nabla f_i(k)(x_k)\|} \nabla f_i(k)(x_k)$

- Output: $\hat{x}_N$, the best among the productive.
Does it converge?

**Theorem [Ju12]:** Let $X$ be bounded and $L$ be the smoothness const. (upper bound). Then,

- $f_0(\hat{x}_N) - f_0(x^*) \leq \frac{LR}{\sqrt{N}}$
- $f_i(\hat{x}_N) \leq \frac{LR}{\sqrt{N}} \forall i$

**Proof Sketch:** Let $f_0(\hat{x}_N) - f_0(x^*) > \frac{LR}{\sqrt{N}}$

- $\sum_{k=1}^{N} (x_k - x^*)^T \nabla f_i(k)(x_k) / \|\nabla f_i(k)(x_k)\| \leq RN$
  - Non-productive: $\frac{R}{\sqrt{N}} (x_k - x^*)^T \nabla f_i(k)(x_k) / \|\nabla f_i(k)(x_k)\| \geq \frac{R^2}{N}$
  - Productive: $\frac{R}{\sqrt{N}} (x_k - x^*)^T \nabla f_i(k)(x_k) / \|\nabla f_i(k)(x_k)\| > \frac{R^2}{N}$
Composite Objective
PROX BASED METHODS
Composite Objectives

\[\min_{w \in \mathbb{R}^n} \Omega(w) + \sum_{i=1}^{m} l(w' \phi(x_i), y_i)\]

Key Idea: Do not approximate non-smooth part
Proximal Gradient Method

\[ x_{k+1} = \text{argmin}_x f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2 + g(x) \]

• If \( g \) is indicator, then same as projected gr.

• If \( g \) is support function: \( g(x) = \max_{y \in S} x^T y \)
  • Assume min-max interchange

\[ x_{k+1} = x_k - s_k \nabla f(x_k) - s_k \Pi_S \left( \frac{1}{s_k} (x_k - s_k \nabla f(x_k)) \right) \]
Proximal Gradient Method

- \( x_{k+1} = \arg\min_x f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2s_k} \|x - x_k\|^2 + g(x) \)

- If \( g \) is indicator, then same as projected gr.

- If \( g \) is support function: \( g(x) = \max_{y \in S} x^T y \)
  - Assume min-max interchange

\[
x_{k+1} = x_k - s_k \nabla f(x_k) - s_k \Pi_S \left( \frac{1}{s_k} (x_k - s_k \nabla f(x_k)) \right)
\]
Rate of Convergence

**Theorem [Ne04]**: If $f$ is smooth with const. $L$, and $s_k = \frac{1}{L'}$, then proximal gradient method generates $x_k$ such that:

$$f(x_k) - f(x^*) \leq \frac{L \|x_0 - x^*\|^2}{2k}.$$ 

- Can be accelerated to $O(1/k^2)$
- Composite same rate as smooth provided proximal oracle exists!
Bibliography


Bibliography


Thanks for listening