OPER 627: Nonlinear Optimization
Lecture 16: Algorithms for optimization over a simple set

Department of Statistical Sciences and Operations Research
Virginia Commonwealth University

Oct 30, 2013

Today's Outline

1. Constrained gradient descent: conditional gradient method (Frank-Wolfe)
2. Gradient projection method: an iterative algorithm for solving constrained optimization over a simple set
3. Proximal point: extended gradient projection

Quiz

1. What is the optimality conditions for constrained optimization problems over a closed convex set
2. What is the optimality conditions for constrained optimization problems over the nonnegative orthant $\mathbb{R}^n_+$?
3. What is a projection? What are the properties of a projection operator?

Constrained gradient descent algorithms

Algorithms in this lecture:

1. They do not rely on any structure of the constraint set other than convexity

2. They generate sequences of feasible points by search along descent directions

Algorithm ingredient:

- Feasible direction $d \neq 0$: x is feasible, if $x + \alpha d$ is feasible for all $\alpha > 0$ that is small enough
- Descent direction d: d is feasible, and $\nabla f(x)^\top d < 0$

Algorithm framework

1. Start at a feasible solution x^0
2. Generate a sequence of feasible solutions $x^{k+1} = x^k + \alpha_k d^k$
   - d^k is a feasible and descent direction
   - α_k is chosen so that $f(x^k + \alpha_k d^k) < f(x^k)$
3. Stepsize rule on α_k:
   - Armijo criterion
   - Constant step size $\alpha_k = 1$

Q: How to choose an initial feasible solution x^0?

A: When C is a polyhedron, i.e., defined by systems of linear equations/inequalities, we can find one by solving a linear program

Conditional gradient method (Frank Wolfe)

Conditional gradient method

A straghtforward way to obtain a descent direction:

$$\min \nabla f(x^k)^\top (x - x^k) \text{ s.t. } x \in C$$

The optimal solution $\bar{x}$, $d^k = \bar{x} - x^k$

- $\bar{x}$ will always be on the boundary of C
- Makes sense only when this problem is much easier to solve than the original problem. E.g., f is nonlinear, C is a polyhedron
- Convergence could be very slow: sublinear convergence, $\lim_{k\to\infty} \frac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|} = 1$ in some cases
- Works well for problems with a low requirement on solution accuracy

Fundamental theorem in gradient projection

Theorem

Ω is a nonempty closed convex set, let $x^* \in \Omega$

(a) $x^* \in \operatorname{argmin}_{x\in\Omega} f(x) \Rightarrow P_\Omega(x^* - \lambda\nabla f(x^*)) = x^*,\ \forall\lambda > 0$

(b) *If $P_\Omega(x^* - \lambda\nabla f(x^*)) = x^*$ for some $\lambda > 0$, and f is convex, then $f(x^*) = \min_{x\in\Omega} f(x)$*

- Note something interesting here: If $P_\Omega(x^* - \lambda\nabla f(x^*)) = x^*$ for some $\lambda > 0$, then $P_\Omega(x^* - \lambda\nabla f(x^*)) = x^*$ for all $\lambda > 0$
- Condition (a) is called gradient projection optimality condition (GPOC)
- GPOC is a generalized definition for stationary point, $\nabla f(x) = 0$

Gradient projection algorithm

1. GPOC can be seen as a fixed point structure: $F(x^*) = x^*$, where $F(x^*) = P_\Omega(x^* - \lambda\nabla f(x^*))$
2. GPOC can be seen as steepest descent + projection, which is intuitive!

Recall: Stepsize selection problem? $\phi(\alpha) = f(x_k + \alpha p_k)$
Wolfe condition

- $\phi(\alpha) \leq \phi(0) + c_1\phi'(0)\alpha, 0 < c_1 < 1$
- $\phi'(\alpha) \geq c_2\phi'(0), 0 < c_2 < 1$

Q: What is the problem? $\phi(\lambda)$ here is nonsmooth! So $\phi'(\lambda)$ is not available! We cannot use Wolfe condition!

Armijo backtracking algorithm

Armijo criterion:

$$f(x_k(\beta^m\lambda)) \leq f(x_k) + c\nabla f(x_k)^\top (x_k(\beta^m\lambda) - x_k)$$

Choose the smallest m that the above holds, $\beta \in (0, 1)$

- Choose an initial λ
- Try points $P_\Omega(x_k - \beta^m\lambda\nabla f(x_k))$, for $m = 0, 1, \ldots$
- Stop when sufficient decrease holds

Theorem

(a) *There always exists a qualifying stepsize that satisfies Armijo criterion*

(b) *Gradient projection algorithm converges to a generalized stationary point*

Rate of convergence

Consider a strictly convex quadratic function $f(x) = \frac{1}{2}x^\top Qx - b^\top x$, let x^* be the unique minimizer of f over Ω. Consider using a constant step size s, then:

$$\begin{aligned}
\|x^{k+1} - x^*\| &= \|[P_\Omega(x^k - s\nabla f(x^k))] - [P_\Omega(x^* - s\nabla f(x^*))]\| \\
&\leq \|(x^k - s\nabla f(x^k)) - (x^* - s\nabla f(x^*))\| \\
&= \|(I - sQ)(x^k - x^*)\| \\
&\leq \max\{|1 - sm|, |1 - sM|\}\|x^k - x^*\|
\end{aligned}$$

where m and M are the smallest and largest eigenvalues of Q.

Concern on gradient projection

1. Convergence rate same as steepest descent, which is slow!
2. Gradient projection is still hard! Projection operator is really heavy
   - Work well on REALLY simple constraints, e.g., box constraints, where the projection is easy

Proximal point

Proximal point

$$\text{prox}_P(x) = \underset{y}{\text{argmin}} \frac{1}{2}\|x - y\|_2^2 + P(y)$$

where:

- $P(y)$ is an extended value convex function: can take value $+\infty$ and $-\infty$
- "$=$" is well-defined because of strong convexity

Examples:

- Indicator function: $\mathbf{1}_C(x) = 0$ if $x \in C$, and $+\infty$ if $x \notin C$
  Q: What is $\text{prox}_{\mathbf{1}_C}(x)$? $\text{proj}_C(x)$!
- $P(x) = \frac{\mu}{2}\|x\|_2^2$
  Q: What is $\text{prox}_P(x)$? $\frac{1}{1+\mu}x$, shrink x towards origin

Proximal point: extended gradient projection

A decomposed unconstrained problem

$$\min \; h(x) = f(x) + P(x)$$

where f is smooth, and P is convex

Extended projection gradient: iterative alternating between proximal point and gradient direction

$$x_{k+1} = \mathrm{prox}_{\alpha_k P}(x_k - \alpha_k \nabla f(x_k))$$

Theorem

If f is convex, P is convex, then $x^* \in \mathrm{argmin}_x \, f(x) + P(x)$ if and only if $x^* = \mathrm{prox}_{vP}(x^* - v\nabla f(x^*))$

Proximal point method

Q: How about solving constrained optimization problem?

$$\min_{x \in C} f(x) \Leftrightarrow \min f(x) + \mathbf{1}_C(x)$$

Advantage of proximal point method:

- Allow us to solve nonsmooth function minimization at a linear rate
- For more information, check out Convex Optimization by Boyd

Next time

Penalty functions for general constrained optimization
Chapter 17 NW book