IE 521: Convex Optimization **Spring 2017, UIUC**

Lecture 24: Subgradient Method – April 24

Instructor: Niao He *Scribe: Shuanglong Wang*

Courtesy warning: These notes do not necessarily cover everything discussed in the class. Please email TA (swang157@illinois.edu) if you find any typos or mistakes.

24.1 Optimization Algorithms and Convergence Rates

Till now, we have discussed several important algorithms for solving convex optimization:

- Ellipsoid method: poly-time algorithm, black-box method, requires first-order and separation oracles
- Interior point method: poly-time algorithm, barrier method, requires structural assumptions on the domain and self-concordant barriers
- Newton method: (local) quadratic convergent algorithm, black-box method, requires smoothness assumptions on the objective and first order and second-order oracles

While the above algorithms have the capability to solve convex programs to high accuracy within a small number of iterations, they suffer from very expensive iteration cost (often cubicly in terms of the problem size), which eventually become impractical for large-scale convex problems.

First-order Methods:

For large-scale convex optimization, simpler algorithms such as first-order methods essentially become the only method of the choices. There exists a dedicated library of efficient first-order optimization algorithms:

- Gradient descent
- Nesterov's accelerated gradient descent and variants (FISTA, geometric descent, etc)
- Coordinate descent and many variants
- Conditional gradient (a.k.a. Frank-Wolfe method)
- Subgradient methods
- Primal-dual methods (Arrow-Hurwicz method, etc)
- Proximal and operator splitting methods (proximal gradient method, ADMM, etc)
- Stochastic and incremental gradient methods (stochastic gradient descent, SVRG, etc)

In the rest of the semester we will present several primary method mainly for the generic non-differentiable constrained problems.

Rate of Convergence:

Suppose the sequence $\{x_k\}$ converges to x^*

Definition 24.1 (Q-convergence) *The convergence rate is said to be*

- *linear: if* $\exists q \in (0,1)$*, such that* $\limsup_{k\to+\infty} \frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|} = q$
- *suplinear: if* $\limsup_{k\to+\infty} \frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|} = 0$
- *sublinear: if* $\limsup_{k\to+\infty} \frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|} = 1$
- *quadratic: if* $\limsup_{k\to+\infty} \frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|^2} < +\infty$

These are also called Q-convergence (quotient).

Example: $x_k = \frac{1}{2^k}, \frac{1}{k}, (\frac{1}{2})^{2^k}$ are Q-linear, Q-sublinear, Q-superlinear convergence.

Note that Q-convergence can be troublesome, for example, consider the sequence

$$x_k = \begin{cases} 1 + \frac{1}{2^k}, & \text{k is even} \\ 1, & \text{k is odd.} \end{cases}$$

Definition 24.2 (R-convergnece) *We say* $\{x_k\}$ *converge to* x^* *R-linearly if* $\exists \{\delta_k\}$ *s.t.* $\| x_k - x^* \| \leq \delta_k$ *and* $\{\delta_k\}$ *converges Q-linearly to* x^*.

For simplicity, we will drop the 'Q' and 'R'.

24.2 Subgradient Method

Consider the generic convex minimization

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in X \end{aligned}$$

where

- f is convex and possibly non-differentiable
- X is non-empty, closed and convex.

Assume the problem is solvable with optimal solution and value denoted as x^*, f^*. Recall that a vector $g \in \partial f(x)$ is called a subgradient of f at x if

$$f(y) \geq f(x) + g^T(y - x) \quad \forall y \in dom(f)$$

The subgradient inequality can be interpreted as a supporting hyperplane for the level set $L_{f(x)}(f) = \{y : f(y) \leq f(x)\}$:

$$g^T(y - x) \leq 0, \forall y \in L_{f(x)}(f) = \{y : f(y) \leq f(x)\}$$

24.2.1 Subgradient Method (N.Shor, 1967)

Subgradient method works as follows: start with $x_1 \in X$ and update

$$x_{t+1} = \Pi_X(x_t - \gamma_t g_t)$$

where

- $g_t \in \partial f(x_t)$ is a subgradient of f at x_t.
- $\gamma_t > 0$ is a proper stepsize
- $\Pi_X(x) = \arg\min_{y \in X} \parallel y - x \parallel_2$ is the Euclidean projection.

Remark:

- When $X = \mathbf{R}^n$, f is continuously differentiable, this reduces to

  $$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

  which is the well-known Gradient Descent method
- Unlike Gradient Descent, the negative direction of subgradient is not always a descent direction.

Choices of Stepsize:

- Constant stepsize: $\gamma_t = \gamma$, $\forall t$
- Diminishing stepsize: $\gamma_t \to 0$ and $\sum_{t=1}^{\infty} \gamma_t = +\infty$, e.g. $\gamma_t = \frac{1}{\sqrt{t}}$
- Square summable stepsize: $\sum_{t=1}^{\infty} \gamma_t^2 < +\infty$ and $\sum_{t=1}^{\infty} \gamma_t = +\infty$, e.g. $\gamma_t = \frac{1}{t}$.
- Scaled stepsize: $\gamma_t = \frac{\gamma}{\|g_t\|_2}$
- Dynamic stepsize (Polyak's optimal stepsize): $\gamma_t = \frac{f(x_t) - f^*}{\|g_t\|_2^2}$