IE 521 Convex Optimization

Lecture 20: Large-Scale Optimization

Subgradient Method

Niao He

1st May 2019

Outline

Algorithms Discussed So Far

- 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
  - Requires solving Newton systems
- Newton Method
  - Local quadratic convergent algorithm
  - Black-box method
  - Requires smoothness assumptions on the objective
  - Requires first-order and second-order oracles

What's in common?

Algorithms Discussed So Far

- 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
  - Requires solving Newton systems
- Newton Method
  - Local quadratic convergent algorithm
  - Black-box method
  - Requires smoothness assumptions on the objective
  - Requires first-order and second-order oracles

High accuracy, but expensive iteration cost. Not scalable!

First-Order Methods

For large-scale convex optimization, simpler algorithms such as first-order methods become the only methods of choice.

- Gradient descent
- Nesterov's accelerated gradient descent and variants
- Coordinate descent and many variants
- Conditional gradient methods
- Subgradient methods
- Primal-dual methods
- Proximal and operator splitting methods
- Stochastic and incremental gradient methods

First-Order Methods

For large-scale convex optimization, simpler algorithms such as first-order methods become the only methods of choice.

- Gradient descent
- Nesterov's accelerated gradient descent and variants
- Coordinate descent and many variants
- Conditional gradient methods
- Subgradient methods
- Primal-dual methods
- Proximal and operator splitting methods
- Stochastic and incremental gradient methods

Moderate accuracy, but cheap iteration cost.

General Constrained Convex Problems

We will focus on the general convex problem:

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

Assumptions

- X is simple and admits easy-to-compute projections
- First-order oracles for $f_0(x), f_i(x)$ are available

Note $f_0(x), f_i(x)$ are not necessarily differentiable or smooth

General Constrained Convex Problems

We will focus on the general convex problem:

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

Assumptions

- X is simple and admits easy-to-compute projections
- First-order oracles for $f_0(x), f_i(x)$ are available

Note $f_0(x), f_i(x)$ are not necessarily differentiable or smooth

What can we do?

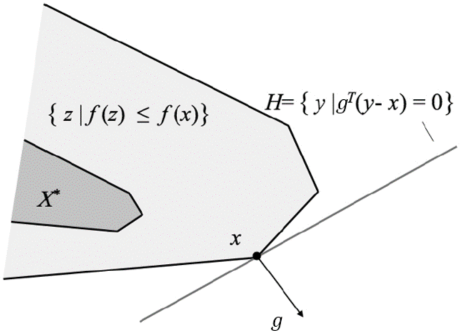
Simple Constrained Convex Problem

Let us start with the simple constrained case:

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

- f is convex and possibly non-differentiable
- X is non-empty, closed and convex
- The problem is solvable with optimal solution and value denoted as x^*, f^*.

Subgradient



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

Subgradient yields a supporting hyperplane for the level set

Subgradient Method (N. Shor, 1967)

---

0. Initialize $x_1 \in X$
1. For $t \geq 1$, do

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

- $g_t \in \partial f(x_t)$ is a subgradient of f at x_t.
- $\gamma_t > 0$ is a proper stepsize
- $\Pi_X(x) = \mathrm{argmin}_{y \in X} \|y - x\|_2$ is the projection.

Remark. When f is differentiable, this reduces to Gradient Descent Method.

Projection

$$\Pi_X(x) = \operatorname*{argmin}_{y \in X} \|y - x\|_2$$

Lemma. $\forall x \in \mathbb{R}^n$, $z \in X$,

$$\|x - z\|_2^2 \geq \|x - \Pi_X(x)\|_2^2 + \|z - \Pi_X(x)\|_2^2$$

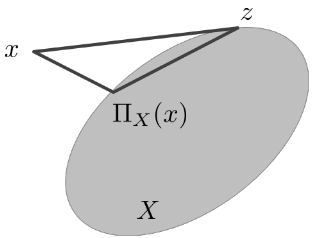


Figure: Projection onto a convex set

Questions

- Is subgradient method a descent method?
- Does it converge?
- How fast does it converge?
- How to choose stepsizes?
- What can we do to improve subgradient method?

Choices of Stepsize

- Constant stepsize:

$$\gamma_t = \gamma$$

- Scaled stepsize:

$$\gamma_t = \frac{\gamma}{\|g_t\|_2}$$

- Non-summable but diminishing stepsize:

$$\gamma_t \to 0 \text{ and } \sum_{t=1}^{\infty} \gamma_t = +\infty$$

- Square summable stepsize:

$$\sum_{t=1}^{\infty} \gamma_t^2 < +\infty \text{ and } \sum_{t=1}^{\infty} \gamma_t = +\infty$$

- Dynamic stepsize:

$$\gamma_t = \frac{f(x_t) - f^*}{\|g_t\|_2^2}$$

Illustration

$$\min_x \quad \|Ax - b\|_1$$

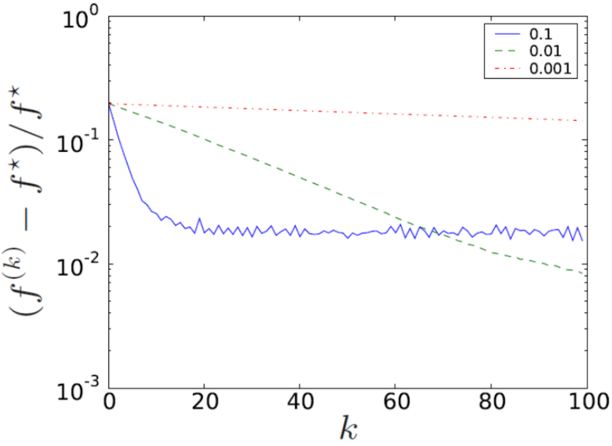


Figure: Fixed Stepsize $\gamma = 0.1, 0.01, 0.001$

Basic "Descent" Lemma

Lemma. We have

$$\|x_{t+1} - x^*\|_2^2 \leq \|x_t - x^*\|_2^2 - 2\gamma_t(f(x_t) - f^*) + \gamma_t^2\|g_t\|_2^2 \quad (\star)$$

Basic "Descent" Lemma

Lemma. We have

$$\|x_{t+1} - x^*\|_2^2 \leq \|x_t - x^*\|_2^2 - 2\gamma_t(f(x_t) - f^*) + \gamma_t^2\|g_t\|_2^2 \quad (\star)$$

Proof.

$$\begin{aligned}\|x_{t+1} - x^*\|_2^2 &= \|\Pi_X(x_t - \gamma_t g_t) - x^*\|_2^2 \\ &\leq \|x_t - \gamma_t g_t - x^*\|_2^2 \\ &= \|x_t - x^*\|_2^2 - 2\gamma_t g_t^T(x_t - x^*) + \gamma_t^2\|g_t\|^2\end{aligned}$$

Basic "Descent" Lemma

Lemma. We have

$$\|x_{t+1} - x^*\|_2^2 \leq \|x_t - x^*\|_2^2 - 2\gamma_t(f(x_t) - f^*) + \gamma_t^2\|g_t\|_2^2 \quad (\star)$$

Proof.

$$\begin{aligned}\|x_{t+1} - x^*\|_2^2 &= \|\Pi_X(x_t - \gamma_t g_t) - x^*\|_2^2 \\ &\leq \|x_t - \gamma_t g_t - x^*\|_2^2 \\ &= \|x_t - x^*\|_2^2 - 2\gamma_t g_t^T(x_t - x^*) + \gamma_t^2\|g_t\|^2\end{aligned}$$

Due to convexity of f, we have $f^* \geq f(x_t) + g_t^T(x^* - x_t)$, i.e.

$$g_t^T(x_t - x^*) \geq f(x_t) - f^*.$$

This leads to $(\star)$.

Polyak's Stepsize

- Minimizing the surrogate function yields the optimal stepsize (Polyak, 1987):

$$\gamma_t = \frac{f(x_t) - f^*}{\|g_t\|_2^2}$$

Polyak's Stepsize

- Minimizing the surrogate function yields the optimal stepsize (Polyak, 1987):

$$\gamma_t = \frac{f(x_t) - f^*}{\|g_t\|_2^2}$$

- This also guarantees strict error reduction:

$$\|x_{t+1} - x_*\|_2^2 \leq \|x_t - x_*\|_2^2 - \frac{(f(x_t) - f_*)^2}{\|g(x_t)\|_2^2}$$

Polyak's Stepsize

- Minimizing the surrogate function yields the optimal stepsize (Polyak, 1987):

$$\gamma_t = \frac{f(x_t) - f^*}{\|g_t\|_2^2}$$

- This also guarantees strict error reduction:

$$\|x_{t+1} - x_*\|_2^2 \leq \|x_t - x_*\|_2^2 - \frac{(f(x_t) - f_*)^2}{\|g(x_t)\|_2^2}$$

- It follows that $f(x_t) \to f^*$ and $\{x_t\} \to x^*$. (why?)

Polyak's Stepsize

- Only useful when f^* is known, e.g., when solving convex feasibility problem:

$$\text{Find } x^* \in X, \text{ s.t. } f_i(x) \leq 0,\ i = 1, ..., m.$$

$$\iff \quad \min_{x \in X} \sum_{i=1}^{m} \max(f_i(x), 0)$$

Polyak's Stepsize

- Only useful when f^* is known, e.g., when solving convex feasibility problem:

$$\text{Find } x^* \in X, \text{ s.t. } f_i(x) \leq 0,\ i = 1, ..., m.$$

$$\iff \quad \min_{x \in X} \sum_{i=1}^{m} \max(f_i(x), 0)$$

- In practice, f^* is often not available. One can replace f^* by an online estimate, e.g.,

$$\hat{f}_t := \min_{0 \leq \tau \leq t} f(x_\tau) - \delta.$$

Main Convergence Result

Theorem. The subgradient method satisfies:

$$\min_{1\le t\le T} f(x_t) - f^* \le \frac{\|x_1 - x^*\|_2^2 + \sum_{t=1}^T \gamma_t^2 \|g_t\|_2^2}{2\sum_{t=1}^T \gamma_t}.$$

Convex Lipschitz Problem

We consider a nice but general problem class:

- $f(x)$ is convex and Lipschitz continuous on X:

$$|f(x) - f(y)| \leq M_f \|x - y\|_2, \quad \forall x, y \in X$$

  where $M_f < +\infty$. (This implies that $\|g_t\|_2 \leq M_f$.)

- X is convex and compact:

$$D_X := \max_{x,y \in X} \|x - y\|_2 < +\infty.$$

Convergence Under Different Stepsizes

- Constant stepsize: $\gamma_t \equiv \gamma$

$$\liminf_{t\to\infty} f(x_t) \le f^* + \frac{M_f^2 \gamma}{2}.$$

Convergence Under Different Stepsizes

- Constant stepsize: $\gamma_t \equiv \gamma$

$$\liminf_{t\to\infty} f(x_t) \leq f^* + \frac{M_f^2\gamma}{2}.$$

- Scaled stepsize: $\gamma_t = \frac{\gamma}{\|g(x_t)\|_2}$

$$\liminf_{t\to\infty} f(x_t) \leq f^* + \frac{M_f\gamma}{2}.$$

Convergence Under Different Stepsizes

- Constant stepsize: $\gamma_t \equiv \gamma$

$$\liminf_{t\to\infty} f(x_t) \le f^* + \frac{M_f^2\gamma}{2}.$$

- Scaled stepsize: $\gamma_t = \frac{\gamma}{\|g(x_t)\|_2}$

$$\liminf_{t\to\infty} f(x_t) \le f^* + \frac{M_f\gamma}{2}.$$

- Non-summable but square-summable stepsize:

$$\liminf_{t\to\infty} f(x_t) = f^*.$$

Convergence Under Different Stepsizes

- Constant stepsize: $\gamma_t \equiv \gamma$

$$\liminf_{t\to\infty} f(x_t) \leq f^* + \frac{M_f^2 \gamma}{2}.$$

- Scaled stepsize: $\gamma_t = \frac{\gamma}{\|g(x_t)\|_2}$

$$\liminf_{t\to\infty} f(x_t) \leq f^* + \frac{M_f \gamma}{2}.$$

- Non-summable but square-summable stepsize:

$$\liminf_{t\to\infty} f(x_t) = f^*.$$

- Non-summable but diminishing stepsize:

$$\liminf_{t\to\infty} f(x_t) = f^*. \quad (\text{why?})$$

Convergence Rate for Convex Lipschitz Problem

Remark.

- In particular, if we set $\gamma_t = \frac{D_X}{M_f\sqrt{t}}$, it holds that

$$\min_{1\leq t\leq T} f(x_t) - f_* \leq O\left(\frac{D_X M_f \ln(T)}{\sqrt{T}}\right).$$

$$\min_{\frac{T}{2}\leq t\leq T} f(x_t) - f_* \leq O\left(\frac{D_X M_f}{\sqrt{T}}\right).$$

Convergence Rate for Convex Lipschitz Problem

Remark.

- In particular, if we set $\gamma_t = \frac{D_X}{M_f\sqrt{t}}$, it holds that

$$\min_{1\leq t\leq T} f(x_t) - f_* \leq O\left(\frac{D_X M_f \ln(T)}{\sqrt{T}}\right).$$

$$\min_{\frac{T}{2}\leq t\leq T} f(x_t) - f_* \leq O\left(\frac{D_X M_f}{\sqrt{T}}\right).$$

- When T is known, setting $\gamma_t \equiv \frac{D_X}{M_f\sqrt{T}}$, we have

$$f(\hat{x}_T) - f^* \leq \frac{D_X M_f}{\sqrt{T}}$$

Convergence Rate for Convex Lipschitz Problem

Remark.

- In particular, if we set $\gamma_t = \frac{D_X}{M_f \sqrt{t}}$, it holds that

$$\min_{1 \le t \le T} f(x_t) - f_* \le O\left(\frac{D_X M_f \ln(T)}{\sqrt{T}}\right).$$

$$\min_{\frac{T}{2} \le t \le T} f(x_t) - f_* \le O\left(\frac{D_X M_f}{\sqrt{T}}\right).$$

- When T is known, setting $\gamma_t \equiv \frac{D_X}{M_f \sqrt{T}}$, we have

$$f(\hat{x}_T) - f^* \le \frac{D_X M_f}{\sqrt{T}}$$

- Subgradient method converges sublinearly. For an accuracy $\epsilon > 0$, need $O(\frac{D_X^2 M_f^2}{\epsilon^2})$ number of iterations.

Strongly Convex and Lipschitz Problem

We now consider an even nicer problem class:

- $f(x)$ is μ-strongly convex on X with $\mu > 0$:

$$f(x) \geq f(y) + \nabla f(y)^T (x - y) + (\mu/2)\|x - y\|_2^2. \quad \forall x, y \in X$$

- $f(x)$ is M_f-Lipschitz continuous on X with $M_f < +\infty$:

$$|f(x) - f(y)| \leq M_f \|x - y\|_2, \quad \forall x, y \in X.$$

Convergence for Strongly Convex Lipschitz Case

Lemma.

$$\|x_{t+1}-x^*\|_2^2 \leq (1-\mu\gamma_t)\|x_t-x^*\|_2^2-2\gamma_t(f(x_t)-f^*)+\gamma_t^2\|g_t\|_2^2 \ (*)$$

Convergence for Strongly Convex Lipschitz Case

Lemma.

$$\|x_{t+1}-x^*\|_2^2 \leq (1-\mu\gamma_t)\|x_t-x^*\|_2^2-2\gamma_t(f(x_t)-f^*)+\gamma_t^2\|g_t\|_2^2 \quad (*)$$

Theorem. Let f be μ-strongly convex and M_f-Lipschitz continuous on X, then with $\gamma_t = \frac{2}{\mu(t+1)}$, we have

$$\min_{1\leq t\leq T} f(x_t) - f_* \leq \frac{2M_f^2}{\mu \cdot (T+1)}.$$

Proof of Convergence

By $(*)$, we have

$$
\begin{aligned}
(f(x_t) - f^*) &\leq \frac{1-\mu\gamma_t}{2\gamma_t}\|x_t - x^*\|_2^2 - \frac{1}{2\gamma_t}\|x_{t+1} - x^*\|_2^2 + \frac{\gamma_t}{2}\|g_t\|_2^2 \\
&= \frac{\mu(t-1)}{4}\|x_t - x^*\|_2^2 - \frac{\mu(t+1)}{4}\|x_{t+1} - x^*\|_2^2 \\
&\quad + \frac{1}{\mu(t+1)}\|g_t\|_2^2
\end{aligned}
$$

Proof of Convergence

By $(*)$, we have

$$
\begin{aligned}
(f(x_t) - f^*) &\leq \frac{1-\mu\gamma_t}{2\gamma_t}\|x_t - x^*\|_2^2 - \frac{1}{2\gamma_t}\|x_{t+1} - x^*\|_2^2 + \frac{\gamma_t}{2}\|g_t\|_2^2 \\
&= \frac{\mu(t-1)}{4}\|x_t - x^*\|_2^2 - \frac{\mu(t+1)}{4}\|x_{t+1} - x^*\|_2^2 \\
&\quad + \frac{1}{\mu(t+1)}\|g_t\|_2^2
\end{aligned}
$$

Hence,

$$
\sum_{t=1}^{T} t(f(x_t) - f^*) \leq -\frac{\mu(T+1)}{4}\|x_{T+1} - x^*\|_2^2 + \frac{T}{\mu}\|g_t\|_2^2
$$

Proof of Convergence

By $(*)$, we have

$$
\begin{aligned}
(f(x_t) - f^*) &\le \frac{1-\mu\gamma_t}{2\gamma_t}\|x_t - x^*\|_2^2 - \frac{1}{2\gamma_t}\|x_{t+1} - x^*\|_2^2 + \frac{\gamma_t}{2}\|g_t\|_2^2 \\
&= \frac{\mu(t-1)}{4}\|x_t - x^*\|_2^2 - \frac{\mu(t+1)}{4}\|x_{t+1} - x^*\|_2^2 \\
&\quad + \frac{1}{\mu(t+1)}\|g_t\|_2^2
\end{aligned}
$$

Hence,

$$
\sum_{t=1}^{T} t(f(x_t) - f^*) \le -\frac{\mu(T+1)}{4}\|x_{T+1} - x^*\|_2^2 + \frac{T}{\mu}\|g_t\|_2^2
$$

$$
\min_{1\le t\le T} f(x_t) - f^* \le \frac{TM_f^2/\mu}{\sum_{t=1}^{T} t} = \frac{2M_f^2}{\mu\cdot(T+1)}
$$

Summary of Subgradient Method

Convex and Lipschitz Continuous Problem

- Stepsize rule: $O(\frac{1}{\sqrt{t}})$
- Convergence rate: $O(\frac{D_X M_f}{\sqrt{t}})$
- Iteration complexity: $O(\frac{D_X^2 M_f^2}{\epsilon^2})$

Strongly Convex and Lipschitz Continuous Problem

- Stepsize rule: $O(\frac{1}{\mu t})$
- Convergence rate: $O(\frac{M_f^2}{\mu t})$
- Iteration complexity: $O(\frac{M_f^2}{\mu \epsilon})$

References

- Nesterov(2004), Chapter 3.2.3, 3.3