Math 273a: Optimization
Gradient descent

Instructor: Wotao Yin
Department of Mathematics, UCLA
Fall 2015

slides based on Chong-Zak, 4th Ed.
online discussions on piazza.com
Main features of gradient methods

- The most popular methods (in continuous optimization)
- Simple and intuitive
- Work under very few assumptions
  (although they cannot directly handle nondifferentiable objectives and constraints, without applying smoothing techniques)
- Work together with many other methods: duality, splitting, coordinate descent, alternating direction, stochastic, online, etc.
- Suitable for large-scale problems, e.g., easy to parallelize for problems with many terms in the objective
Gradients

- We let $\nabla f(x_0)$ denote the gradient of $f$ at point $x_0$.
- $\nabla f(x_0) \perp$ tangent of the levelset curve of $f$ passing $x_0$, pointing outward (recall: level set $\mathcal{L}_f(c) := \{x : f(x) = c\}$)
\( \nabla f(x_0) \) is **max-rate ascending direction** of \( f \) at \( x_0 \) (for a small displacement), and \( \| \nabla f(x_0) \| \) is the rate.

**Reason:** pick any direction \( d \) with \( \| d \| = 1 \). The rate of change at \( x \) is

\[
\langle \nabla f(x), d \rangle \leq \| \nabla f(x) \| \cdot \| d \| = \| \nabla f(x) \|.
\]

If we set \( d = \nabla f(x)/\| \nabla f(x) \| \), then

\[
\langle \nabla f(x), d \rangle = \| \nabla f(x) \|.
\]

Therefore, \(-\nabla f(x)\) is the **max-rate descending direction** of \( f \) and a **good search direction**.
A negative gradient step can decrease the objective

- Let $x^{(0)}$ be any initial point.
- First-order Taylor expansion for candidate point $x = x^{(0)} - \alpha \nabla f(x^{(0)})$:
  \[
f(x) - f(x^{(0)}) = -\alpha \|\nabla f(x^{(0)})\|^2 + o(\alpha)
  \]
- Hence, if $\nabla f(x^{(0)}) \neq 0$ (the first-order necessary condition is not met) and $\alpha$ is sufficiently small, we have
  \[
f(x) < f(x^{(0)}).
  \]
- Therefore, for sufficiently small $\alpha$, $x$ is an improvement over $x^{(0)}$. 
Gradient descent algorithm

- Given initial $x^{(0)}$, the gradient descent algorithm uses the following update to generate $x^{(1)}$, $x^{(2)}$, ..., until a stopping condition is met:
  from the current point $x^{(k)}$, generate the next point $x^{(k+1)}$ by
  $$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}),$$
- $\alpha_k$ is called the step size
Alternative interpretation:

- notice that

\[ x^{(k+1)} = \arg \min_x \frac{1}{2\alpha_k} \left\| x - (x^{(k)} - \alpha_k \nabla f(x^{(k)})) \right\|^2 \]

\[ = \arg \min_x f(x^{(k)}) + \langle \nabla f(x^{(k)}), x - x^{(k)} \rangle + \frac{1}{2\alpha_k} \| x - x^{(k)} \|^2 \]

(2nd “=” follows from that adding constants and multiplying a positive constant do not change the set of minimizers or “arg min”)

- Hence, \( x^{(k+1)} \) is obtained by minimizing the linearization of \( f \) at \( x^{(k)} \) and a proximal term that keeps \( x^{k+1} \) close to \( x^{(k)} \).

- The reformulation is useful to develop the extensions of gradient descent:
  - projected gradient method
  - proximal-gradient method
  - accelerated gradient method
  - ......
When to stop the iteration

The first-order necessary condition $\|\nabla f(x^{(k+1)})\| = 0$ is not practical.

Practical conditions:

- gradient condition $\|\nabla f(x^{(k+1)})\| < \epsilon$
- successive objective condition $|f(x^{(k+1)}) - f(x^{(k)})| < \epsilon$ or the relative one
  $$\frac{|f(x^{(k+1)}) - f(x^{(k)})|}{|f(x^{(k)})|} < \epsilon$$
- successive point difference $\|x^{(k+1)} - x^{(k)}\| < \epsilon$ or the relative one
  $$\frac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k)}\|} < \epsilon$$
- to avoid division by tiny numbers (unstable division), we can replace the denominators by $\max\{1, |f(x^{(k)})|\}$ and $\max\{1, \|x^{(k)}\|\}$, respectively
Small versus large step sizes $\alpha_k$

**Small step size:**
- Pros: iterations are more likely converge, closely traces max-rate descends
- Cons: need more iterations and thus evaluations of $\nabla f$

**Large step size:**
- Pros: better use of each $\nabla f(x^{(k)})$, may reduce the total iterations
- Cons: can cause overshooting and zig-zags, too large $\Rightarrow$ diverged iterations
Small versus large step sizes $\alpha_k$

**Small step size:**

- Pros: iterations are more likely converge, closely traces max-rate descends
- Cons: need more iterations and thus evaluations of $\nabla f$

**Large step size:**

- Pros: better use of each $\nabla f(x^{(k)})$, may reduce the total iterations
- Cons: can cause overshooting and zig-zags, too large $\Rightarrow$ diverged iterations

In practice, step sizes are often chosen

- as a fixed value if $\nabla f$ is Lipschitz (rate of change is bounded) with the constant known or an upper bound of it known
- by line search
- by a method called Barzilai-Borwein with nonmonotone line search
Steepest descent method
(gradient descent with exact line search)

Step size $\alpha_k$ is determined by exact minimization

$$
\alpha_k = \arg \min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).
$$

It is used mostly for quadratic programs (with $\alpha_k$ in a closed form) and some problems with inexpensive evaluation values but expensive gradient evaluation; otherwise it is not worth the effort to solve this subproblem exactly.
Proposition 8.1 If \( \{x^{(k)}\}_{k=0}^\infty \) is a steepest descent sequence for a given function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \), then for each \( k \) the vector \( x^{(k+1)} - x^{(k)} \) is orthogonal to the vector \( x^{(k+2)} - x^{(k+1)} \).

□
Steepest descent for quadratic programming

Assume that $Q$ is symmetric and positive definite ($x^TQx > 0$ for any $x \neq 0$).

Consider the quadratic program

$$f(x) = \frac{1}{2}x^TQx - b^Tx$$

with

$$\nabla f(x) = Qx - b.$$

**Steepest descent iteration:** start from any $x^{(0)}$, set

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}, \quad k = 0, 1, 2, \ldots$$

where $g^{(k)} := \nabla f(x^{(k)})$ and

$$\alpha_k = \arg \min_{\alpha \geq 0} f(x^{(k)} - \alpha g^{(k)})$$

$$= \frac{g^{(k)T}g^{(k)}}{g^{(k)T}Qg^{(k)}}.$$
Examples

Example 1: $f(x) = x_1^2 + x_2^2$. Steepest descent arrives at $x^* = 0$ in 1 iteration.

Example 1: $f(x) = \frac{1}{5} x_1^2 + x_2^2$. Steepest descent makes progress in a narrow valley.
Performance of steepest descent

- **Per-iteration cost:** dominated by *two* matrix-vector multiplications:
  - \( g^{(k)} = Qx^{(k)} - b \)
  - computing \( \alpha_k \) involves \( Qg^{(k)} \)

but they can be easily reduced to *one* matrix-vector multiplication.

- **Convergence speed:** determined by the initial point and the spectral condition of \( Q \). To analyze them, we
  - define solution error: \( e^{(k)} = x^{(k)} - x^* \) (not known, an analysis tool)
  - have property: \( g^{(k)} = Qx^{(k)} - b = Qe^{(k)} \).
Good cases:

- $e^{(k)}$ is an eigenvector of $Q$ with eigenvalue $\lambda$

$$
e^{(k+1)} = e_k - \alpha_k g^{(k)} = e^{(k)} - \frac{g^{(k)} g^{(k)}}{g^{(k)T} Q g^{(k)}} (Q e^{(k)})$$

$$= e^{(k)} + \frac{g^{(k)T} g^{(k)}}{\lambda g^{(k)T} g^{(k)}} (-\lambda e^{(k)}) = 0.$$

- $Q$ has only one distinct eigenvalue (the level sets of $Q$ are circles)

The general case: define $\|e\|_A := \sqrt{e^T A e}$ and $\kappa := \lambda_{\text{max}}(Q)/\lambda_{\text{min}}(Q)$, then we have

$$\|e^{(k)}\|_A \leq \left( \frac{\kappa - 1}{\kappa + 1} \right)^k \|e^{(0)}\|_A.$$
A example from *An Introduction to CG method* by Shewchuk
Gradient descent with fixed step size

- Iteration:
  \[ x^{(k+1)} = x^{(k)} - \alpha g^{(k)} \]

- We assume that \( x^* \) exists

- Check distance to solution:
  \[
  \| x^{(k+1)} - x^* \|^2 = \| x^{(k)} - x^* - \alpha g^{(k)} \|^2 \\
  = \| x^{(k)} - x^* \|^2 - 2\alpha \langle g^{(k)}, x^{(k)} - x^* \rangle + \alpha^2 \| g^{(k)} \|^2.
  \]

- Therefore, in order to have \( \| x^{(k+1)} - x^* \| \leq \| x^{(k)} - x^* \| \), we must have
  \[
  \frac{\alpha}{2} \| g^{(k)} \|^2 \leq \langle g^{(k)}, x^{(k)} - x^* \rangle.
  \]

Since \( g^* := \nabla f(x^*) = 0 \), the condition is equivalent to
\[
\frac{\alpha}{2} \| g^{(k)} - g^* \|^2 \leq \langle g^{(k)} - g^*, x^{(k)} - x^* \rangle.
\]
Special case: convex and Lipschitz differentiable $f$

- **Definition:** A function $f$ is *L*-Lipschitz differentiable, $L \geq 0$, if $f \in C^1$ and

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$

(the maximum rate of change of $\nabla f$ is $L$)

- **Baiillon-Haddad theorem:** if $f \in C^1$ is a convex function, then it is *L*-Lipschitz differentiable if and only if

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq L\langle \nabla f(x) - \nabla f(y), x - y \rangle.$$  

(such $\nabla f$ is called $1/L$-cocoercive)
Theorem: Let $f \in C^1$ be a convex function and $L$-Lipschitz differentiable. If $0 < \alpha \leq 2/L$, then
\[
\frac{\alpha}{2} \|g^{(k)} - g^*\|^2 \leq \langle g^{(k)} - g^*, x^{(k)} - x^* \rangle
\]
and thus $\|x^{(k+1)} - x^*\| \leq \|x^{(k)} - x^*\|$ for $k = 0, 1 \ldots$. The iteration stays bounded.

Theorem: Let $f \in C^1$ be a convex function and $L$-Lipschitz differentiable. If $0 < \alpha < L/2$, then
\begin{itemize}
  \item both $f(x^{(k)})$ and $\|\nabla f(x^{(k)})\|$ are monotonically decreasing,
  \item $f(x^{(k)}) - f(x^*) = O\left(\frac{1}{k}\right)$,
  \item $\|\nabla f(x^{(k)})\| = o\left(\frac{1}{k}\right)$.
\end{itemize}
(one often writes $\|\nabla f(x^{(k)})\|^2 = o\left(\frac{1}{k^2}\right)$ since $\|\nabla f(x^{(k)})\|^2$ naturally appears in most analysis.)
Gradient descent with fixed step size
for quadratic programming

Assume that $Q$ is symmetric and positive definite ($x^T Q x > 0$ for any $x \neq 0$).

Consider the quadratic program

$$f(x) = \frac{1}{2} x^T Q x - b^T x$$

**Theorem 8.3** For the fixed-step-size gradient algorithm, $x^{(k)} \to x^*$ for any $x^{(0)}$ if and only if

$$0 < \alpha < \frac{2}{\lambda_{\text{max}}(Q)}.$$
Summary

- Negative gradient $-\nabla f(x^{(k)})$ is the max-rate descending direction
- For some small $\alpha_k$, $x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$ improves over $x^{(k)}$
- There are practical rules to determine when to stop the iteration
- Exact line search works for quadratic program with $Q > 0$. Zig-zag occurs if $x^{(0)} - x^*$ is away from an eigenvector and spectrum of $Q$ is spread
- Fixed step gradient descent works for convex and Lipschitz-differentiable $f$
- To keep the discussion short and informative, we have omitted much other convergence analysis.