Math 164: Optimization
Barzilai-Borwein Method

Instructor: Wotao Yin
Department of Mathematics, UCLA
Spring 2015

online discussions on piazza.com
The BB method was published in a 8-page paper\textsuperscript{1} in 1988.

It is a gradient method with modified step sizes, which are motivated by Newton’s method but not involves any Hessian.

At nearly no extra cost, the method often significantly improves the performance of a standard gradient method.

The method is used along with non-monotone line search as a safeguard.

Motivation of the BB method

Let $g^{(k)} = \nabla f(x^{(k)})$ and $F^{(k)} = \nabla^2 f(x^{(k)})$.

- **gradient method:** $x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$
  - choice of $\alpha_k$: fixed, exact line search, or fixed initial + line search
  - **pros:** simple
  - **cons:** no use of 2nd order information, sometimes zig-zag

- **Newton’s method:** $x^{(k+1)} = x^{(k)} - (F^{(k)})^{-1} g^{(k)}$
  - **pros:** 2nd-order information, 1-step for quadratic function, fast convergence near solution
  - **cons:** forming and computing $(F^{(k)})^{-1}$ is expensive, need modifications if $F^{(k)} \neq 0$

The BB method chooses $\alpha_k$ so that $\alpha_k g^{(k)}$ approximates $(F^{(k)})^{-1} g^{(k)}$ without computing $F^{(k)}$
Derive the BB method

- Consider
  \[
  \min_x f(x) = \frac{1}{2} x^T A x - b^T x,
  \]
  where \( A \succ 0 \) is symmetric. Gradient is \( g^{(k)} = A x^{(k)} - b \). Hessian is \( A \).

- Newton step: \( d^{(k)}_{\text{newton}} = -A^{-1} g^{(k)} \)

- **Goal:** choose \( \alpha_k \) so that \( -\alpha_k g^{(k)} = -(\alpha_k^{-1} I)^{-1} g^{(k)} \) approximates \( -A^{-1} g^{(k)} \)

- Define: \( s^{(k-1)} := x^{(k)} - x^{(k-1)} \) and \( y^{(k-1)} := g^{(k)} - g^{(k-1)} \). Then \( A \) satisfies:
  \[
  A s^{(k-1)} = y^{(k-1)}.
  \]

- Therefore, given \( s^{(k-1)} \) and \( y^{(k-1)} \), how about choose \( \alpha_k \) so that
  \[
  (\alpha_k^{-1} I) s^{(k-1)} \approx y^{(k-1)}
  \]
• **Goal:**

\[(\alpha_k^{-1} I) s^{(k-1)} \approx y^{(k-1)}.\]

• **BB method:**

  • Least-squares problem: (let \(\beta = \alpha^{-1}\))

\[
\alpha_k^{-1} = \arg \min_{\beta} \frac{1}{2} \| s^{(k-1)} \beta - y^{(k-1)} \|^2 \quad \Longrightarrow \quad \alpha_k^1 = \frac{(s^{(k-1)})^T s^{(k-1)}}{(s^{(k-1)})^T y^{(k-1)}}
\]

  • Alternative Least-squares problem:

\[
\alpha_k = \arg \min_{\alpha} \frac{1}{2} \| s^{(k-1)} - y^{(k-1)} \alpha \|^2 \quad \Longrightarrow \quad \alpha_k^2 = \frac{(s^{(k-1)})^T y^{(k-1)}}{(y^{(k-1)})^T y^{(k-1)}}
\]

• \(\alpha_k^1\) and \(\alpha_k^2\) are called the BB step sizes.
Apply the BB method

- Since \(x^{(k-1)}\) and \(g^{(k-1)}\) and thus \(s^{(k-1)}\) and \(y^{(k-1)}\) are unavailable at \(k = 0\), we apply the standard gradient descent at \(k = 0\) and start BB at \(k = 1\).

- We can use either \(\alpha^1_k\) or \(\alpha^2_k\) or alternate between them.

- We can fix \(\alpha_k = \alpha^1_k\) or \(\alpha_k = \alpha^2_k\) for a few consecutive steps.

- It performs very well on minimizing quadratic and many other functions.

- However, \(f_k\) and \(\|\nabla f_k\|\) are **not** monotonic!
Steepest descent versus BB on quadratic programming

- **Model:**

  \[
  \text{minimize } f(x) := \frac{1}{2} x^T Ax - b^T x.
  \]

- **Gradient iteration**

  \[
  x^{k+1} \leftarrow x^{(k)} - \alpha_k (Ax^{(k)} - b).
  \]

- **Steepest descent** selects \( \alpha_k \) as \( \arg \min_\alpha f(x^{(k)} - \alpha_k (Ax^{(k)} - b)) \)

  \[
  \alpha_k = \frac{(r^{(k)})^T r^{(k)}}{(r^{(k)})^T A r^{(k)}}
  \]

  where \( r^{(k)} := b - Ax^{(k)} \).

- **BB** selects \( \alpha_k \) as

  \[
  \alpha_k^1 = \frac{(s^{(k-1)})^T s^{(k-1)}}{(s^{(k-1)})^T y^{(k-1)}}
  \]
Numerical example

- Set symmetric matrix $A$ to have the condition number $\frac{\lambda_{\text{max}}(A)}{\lambda_{\text{min}}(A)} = 50$.

- Stopping criterion:
  \[ \| r^{(k)} \| < 10^{-8} \]

- Steepest descent stops in 90 iterations

- BB stops in 10 iterations
Properties of Barzilai-Borwein

- For quadratic functions, it has R-linear convergence\(^2\)
- For 2D quadratic function, it has Q-superlinear convergence\(^3\)
- No convergence guarantee for smooth convex problems. On these problems, we pair up BB with **non-monotone line search**.

\[
\text{BB on } \text{Laplace2: } \min \frac{1}{2} x^T A x - b^T x + \frac{h^2}{4} \sum_{ijk} u^4_{ijk}.
\]

\(^2\)Dai and Liao [2002]  
\(^3\)Barzilai and Borwein [1988], Dai [2013]
Nonmonotone line search

- Some growth in the function value is permitted
- Sometimes improve the likelihood of finding a global optimum
- Improve convergence speed when a monotone scheme is forced to creep along the bottom of a narrow curved valley
- Early nonmonotone line search method developed for Newton’s methods

\[
f(x^{(k)} + \alpha d^{(k)}) \leq \max_{0 \leq j \leq m_k} f(x^{k-j}) + c_1 \alpha \nabla f_k^T d^{(k)}
\]

However, it may still kill R-linear convergence. **Example:** \( x \in \mathbb{R}, \)

minimize \( f(x) = \frac{1}{2} x^2, \quad x^0 \neq 0, \quad d^{(k)} = -x^{(k)}. \)

\[
\alpha_k = \begin{cases} 
1 - 2^{-k}, & k = i^2 \text{ for some integer } i, \\
2, & \text{otherwise},
\end{cases}
\]

converges R-linear but fails to satisfy the condition for \( k \) large.

---

\(^4\)Grippo, Lampariello, and Lucidi [1986]
Zhang-Hager nonmonotone line search\(^5\)

1. initialize \(0 < c_1 < c_2 < 1\), \(C_0 \leftarrow f(x^0), Q_0 \leftarrow 1, \eta < 1, k \leftarrow 0\)

2. while not converged do

3a. compute \(\alpha_k\) satisfying the modified Wolfe conditions OR

3b. find \(\alpha_k\) by backtracking, to satisfy the modified Armijo condition:

4. \(x^{k+1} \leftarrow x^{(k)} + \alpha_k d^{(k)}\)

5. \(Q_{k+1} \leftarrow \eta Q_k + 1, C_{k+1} \leftarrow (\eta Q_k C_k + f(x^{k+1}))/Q_{k+1}\).

Comments:

- If \(\eta = 1\), then \(C_k = \frac{1}{k+1} \sum_{j=0}^{k} f_j\).
- Since \(\eta < 1\), \(C_k\) is a weighted sum of all past \(f_j\), more weights on recent \(f_j\).

---

\(^5\)Zhang and Hager [2004]
Convergence (advanced topic)

The results below are left to the reader as an exercise.

If $f \in C^1$ and bounded below, $\nabla f_k^T d^{(k)} < 0$, then

- $f_k \leq C_k \leq \frac{1}{k+1} \sum_{j=0}^{(k)} f_j$
- there exists $\alpha_k$ satisfying the modified Wolfe or Armijo conditions

In addition, if $\nabla f$ is Lipschitz with constant $L$, then

- $\alpha_k > C \frac{|\nabla f_k^T d^{(k)}|}{\|d^{(k)}\|}$ for some constant depending on $c_1, c_2, L$ and the backing factor

Furthermore, if for all sufficiently large $k$, we have uniform bounds

$$\nabla f_k^T d^{(k)} \leq -c_3 \|\nabla f_k\|^2 \quad \text{and} \quad \|d^{(k)}\| \leq c_4 \|\nabla f_k\|$$

then

- $\lim_{k \to \infty} \nabla f_k = 0$

Once again, pairing with non-monotone linear search, Barzilai-Borwein gradient methods work very well on general unconstrained differentiable problems.
References:


