Barzilai-Borwein Method

Ruishan Lin

STAT 742 Course Project

10/21/2022

Overview

1. Background
   - Optimization Context
   - Gradient Descent
   - Newton's Method
2. BB Method
3. Examples
   - Two-component Gaussian Mixture
   - Logistic Regression

Optimization Context

- Problem

$$\underset{x \in \mathbb{R}^n}{min} \quad f(x)$$

where $f(\cdot)$ is at least twice continuously differentiable and convex.

- 2^{nd} Order Convexity Condition

$$\nabla^2 f(x) \succeq 0 \quad \forall x \in \mathbb{R}^n$$

- Solution

$$x^* \in \underset{x}{argmin} \quad f(x)$$

- In other words, we only consider a smooth, convex, unconstrained optimization problem.

Gradient Descent

Procedure:

1. Initialize $k = 0$ and set $x^{(k)} \leftarrow x_0$ for an initial solution x_0
2. Set $x^{(k+1)} \leftarrow x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)})$. $\alpha^{(k)}$ is the step size.
3. Iterate 2. until a stopping criterion is satisfied.

Newton's Method

Procedure:

1. Initialize $k = 0$ and set $x^{(k)} \leftarrow x_0$ for an initial solution x_0
2. Set $x^{(k+1)} \leftarrow x^{(k)} - [\nabla^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)})$
3. Iterate 2. until a stopping criterion is satisfied.

Barzilai-Borwein Method

- Idea: Find $\alpha^{(k)}$ such that $\alpha^{(k)}\nabla f(x^{(k)})$ approximates $[\nabla^2 f(x^{(k)})]^{-1}\nabla f(x^{(k)})$

- Deriving $\alpha^{(k)}$:
  Let

$$
\begin{aligned}
A &= \nabla^2 f(x^{(k)}) \\
s^{(k-1)} &= x^{(k)} - x^{(k-1)} \\
y^{(k-1)} &= \nabla f(x^{(k)}) - \nabla f(x^{(k-1)})
\end{aligned}
$$

  Then we have

$$A\, s^{(k-1)} = y^{(k-1)} \quad \text{or} \quad s^{(k-1)} = A^{-1} y^{(k-1)}$$

BB Method Cont.

$$[\alpha^{(k)}]^{-1} s^{(k-1)} \approx y^{(k-1)} \quad \text{or} \quad s^{(k-1)} \approx \alpha^{(k)} y^{(k-1)}$$

Apply Least Squares Minimization:

$$[\alpha^{(k)}]^{-1} = \underset{\beta}{argmin} \, ||\beta s^{(k-1)} - y^{(k-1)}||^2$$

or

$$\alpha^{(k)} = \underset{\beta}{argmin} \, ||s^{(k-1)} - \beta y^{(k-1)}||^2$$

$$\Longrightarrow \bar{\beta}^* = \frac{< s^{(k-1)}, y^{(k-1)} >}{||s^{(k-1)}||^2} \quad \tilde{\beta}^* = \frac{< s^{(k-1)}, y^{(k-1)} >}{||y^{(k-1)}||^2}$$

BB Method Cont.

Procedure:

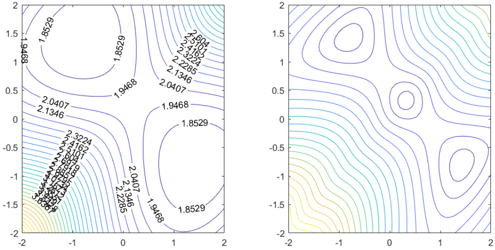
1. Initialize $k = 0$ and set $x^{(k)} \leftarrow x_0$ for an initial solution x_0.
2. Set $x^{(k+1)} \leftarrow x^{(k)} - \alpha^{(k)} \nabla f(x^{(k)})$ where $\alpha^{(k)}$ is either $1/\bar{\beta}^*$ or $\tilde{\beta}^*$.
3. Iterate 2. until a stopping criterion is satisfied.

Two-Component Gaussian Mixture

- With negative Log-likelihood given by

$$f(\mu_1, \mu_2) = -\frac{1}{n}\sum_{i=1}^{n} log(0.5\phi(X_i - \mu_1) + 0.5\phi(X_i - \mu_2))$$

  given data $\{X_i\}_{i=1}^{n}$ and $\mu_1, \mu_2 \in \mathbb{R}$. $n = 50$
- Contour plots of the $f(\cdot)$ and the $||\nabla f(\cdot)||$



Gradient

$$\nabla_{\mu_1} f(\mu_1, \mu_2) = \frac{1}{n} \sum_{i=1}^{n} \frac{\phi(X_i - \mu_1)(X_i - \mu_1)}{\phi(X_i - \mu_1) + \phi(X_i - \mu_2)}$$

$$\nabla_{\mu_2} f(\mu_1, \mu_2) = \frac{1}{n} \sum_{i=1}^{n} \frac{\phi(X_i - \mu_2)(X_i - \mu_2)}{\phi(X_i - \mu_1) + \phi(X_i - \mu_2)}$$

Hessian

$$\nabla^2_{\mu_1^2} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\phi(X_i-\mu_1)(X_i-\mu_1) - \phi(X_i-\mu_1)\phi(X_i-\mu_1)}{[\phi(X_i-\mu_1)+\phi(X_i-\mu_2)]^2}$$

$$\nabla^2_{\mu_2^2} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\phi(X_i-\mu_2)(X_i-\mu_2) - \phi(X_i-\mu_2)\phi(X_i-\mu_2)}{[\phi(X_i-\mu_1)+\phi(X_i-\mu_2)]^2}$$

$$\nabla^2_{\mu_1,\mu_2} = \frac{1}{n}\sum_{i=1}^{n} \frac{\phi(X_i-\mu_1)(X_i-\mu_1)\phi(X_i-\mu_2)(X_i-\mu_2)}{[\phi(X_i-\mu_1)+\phi(X_i-\mu_2)]^2} = \nabla^2_{\mu_2,\mu_1}$$

Starting value $(1, -1)$

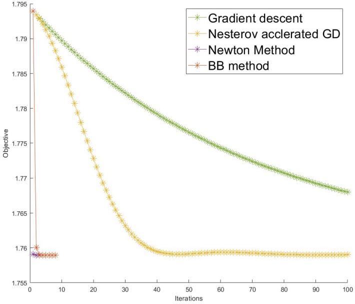


Figure: Plot of objective function against iteration.

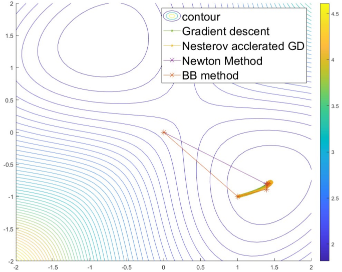


Figure: Contour Plot

Other starting values

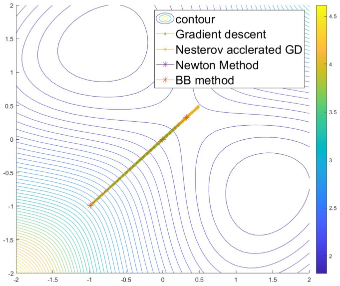


Figure: $(-1, -1)$

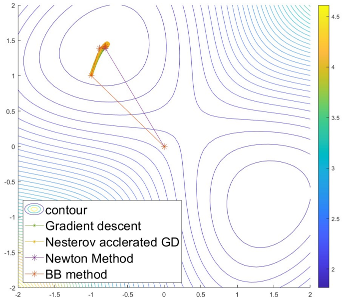


Figure: $(0, 0)$

Second Example: Logistic Regression

$$\min_{\beta\in\mathbb{R},\beta_0\in\mathbb{R}} f(\beta_0,\beta) = \sum_{i=1}^{n} log(1+exp(-y_i(x_i^T\beta+\beta_0)))$$

given data $\{x_i, y_i\}_{i=1}^{n}$

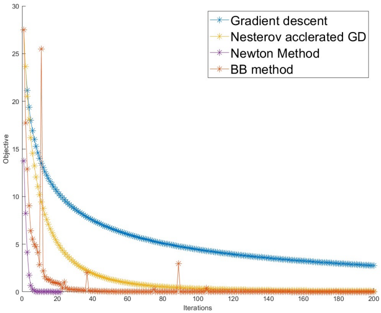
Gradient and Hessian

$$\nabla_{\beta_0} f(\beta_0, \beta) = -\sum_{i=1}^{n} \frac{y_i}{1 + exp(y_i(x_i^T \beta + \beta_0))}$$

$$\nabla_{\beta} f(\beta_0, \beta) = -\sum_{i=1}^{n} \frac{y_i x_i}{1 + exp(y_i(x_i^T \beta + \beta_0))}$$

$$\nabla^2 f(\beta_0, \beta) = \sum_{i=1}^{n} \begin{bmatrix} X \\ 1 \end{bmatrix} \frac{exp(-y_i(x_i^T \beta + \beta_0))}{[1 + exp(-y_i(x_i^T \beta + \beta_0))]^2} \begin{bmatrix} X \\ 1 \end{bmatrix}^T$$

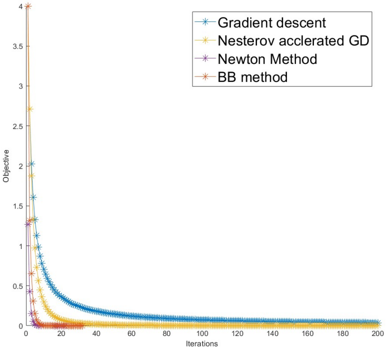
Case when $d = 20, n = 50$



	CPU run time	iterations	gradient norm
Gradient Descent	0	200	0.7986
Accelerated GD	0	200	0.0233
Newton's Method	0.0156	22	6.7168e-09
BB Method	0.0313	200	1.1487e-08

Table: CPU run time,iterations, and optimality

Case when $d = 500, n = 10$



	CPU run time	iterations	gradient norm
Gradient Descent	0.0156	200	0.2487
Accelerated GD	0	200	0.0050
Newton's Method	0.6875	22	5.8615e-09
BB Method	0	32	7.6061e-09

References

1. Slawski, M. (2022, October) *Gradient Descent, Extensions to Gradient Descent* [Slides] Department of Statistics, George Mason University.
2. Yin W. (2015) *Optimization Barzilai Borwein Method* [Slides] Department of Mathematics, UCLA.