E1 260 OptML: Project 2
Barzilai-Borwein Step Size for Stochastic Gradient Descent

A Suresh Varma
MTech SP
21848
varmaa@iisc.ac.in

I BV Lakshmana Raju
MTech CSA
21190
bhargavav@iisc.ac.in

1. Abstarct

The challenge of choosing an appropriate step size in stochastic gradient descent (SGD) methods has led to either using a diminishing step size or time-consuming manual tuning. A new approach, using the Barzilai-Borwein (BB) method, has been proposed to automatically compute step sizes for both SGD and its variant, SVRG. The resulting algorithms, SGD-BB and SVRG-BB, offer a more efficient solution to the step size problem in SGD methods.

2. Introduction

The total number of gradient evaluations in SGD is affected by the variance of stochastic gradients, resulting in sublinear convergence rate for strongly convex and smooth problems. To address this, various variants of SGD have been developed to reduce variance and improve complexity, including SAG, SAGA, SDCA, and SVRG. However, an important issue regarding stochastic algorithms that remains unresolved is how to choose an appropriate step size t. Traditional line search techniques cannot be used due to the limited sub-sampled information available in SGD. Recent works such as AdaGrad, SAG line search, Gaussian process-based line search, and online SGD for step size parameter estimation have been suggested to tackle this problem.

The paper begins with a brief introduction to the deterministic BB method in Section 3. Section 4.1 presents the SGD-BB method with linear convergence for strongly convex functions. A smoothing technique is proposed to improve its performance in Section 4.2. Section 5 contains numerical experiments for both SVRG-BB and SGD-BB methods.

Contributions : A.Suresh Varma for SGD-BB and I.B.V. Lakshmana Raju for SVRG-BB.

3. The Barzilai-Borwein Step Size

The BB method, proposed in [2], is a successful technique for solving nonlinear optimization problems. It is inspired by quasi-Newton methods and is used to solve unconstrained minimization problems.

$$min_x = f(x) \tag{1}$$

where f is differentiable. A typical iteration of quasi-Newton methods for solving (4) is:

$$x_{t+1} = x_t - B_t^{-1}\nabla f(x_t) \quad (5) \tag{2}$$

B_t approximates the Hessian matrix of f at x_t and satisfies the secant equation: $B_t s_t = y_t$, where $s_t = x_t - x_{t-1}$ and $y_t = \nabla f(x_t) - \nabla f(x_{t-1})$ for $t \geq 1$. Solving the linear system in (5) can be time-consuming when B_t is large and dense.To reduce computational burden, BB method replaces B_t with $(1/\eta_t)I$. But a scalar η_t cannot satisfy the secant equation with $B_t = (1/\eta_t)I$. Thus, η_t is obtained by minimizing $|(1/\eta_t)s_t - y_t|_2^2$, leading to the choice of η_t as:

$$\eta_t = \frac{|s_t|_2^2}{s_t^T y_t}. \tag{3}$$

Therefore, a typical iteration of the BB method for solving (1) is

$$x_{t+1} = x_t - \eta_t \nabla f(x_t), \tag{4}$$

where η_t is computed by (3).

4. Barzilai-Borwein Step Size

4.1. Barzilai-Borwein Step Size for SGD

Algorithm 2 SGD with BB step size (SGD-BB)

Parameters: update frequency m, initial step sizes η_0 and η_1 (only used in the first two epochs), weighting parameter $\beta \in (0, 1)$, initial point $\tilde{x}_0$
for $k = 0, 1, \cdots$ **do**
  if $k > 0$ **then**
    $\eta_k = \frac{1}{m} \cdot \|\tilde{x}_k - \tilde{x}_{k-1}\|_2^2 / |(\tilde{x}_k - \tilde{x}_{k-1})^\top (\hat{g}_k - \hat{g}_{k-1})|$
  end if
  $x_0 = \tilde{x}_k$
  $\hat{g}_{k+1} = 0$
  for $t = 0, \cdots, m-1$ **do**
    Randomly pick $i_t \in \{1, \ldots, n\}$
    $x_{t+1} = x_t - \eta_k \nabla f_{i_t}(x_t)$
    $\hat{g}_{k+1} = \beta \nabla f_{i_t}(x_t) + (1-\beta)\hat{g}_{k+1}$
  end for
  $\tilde{x}_{k+1} = x_m$
end for

4.2. Barzilai-Borwein Step Size for SVRG

Algorithm 1 SVRG with BB step size (SVRG-BB)

Parameters: update frequency m, initial point $\tilde{x}_0$, initial step size η_0 (only used in the first epoch)
for $k = 0, 1, \cdots$ **do**
  $g_k = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\tilde{x}_k)$
  if $k > 0$ **then**
    $\eta_k = \frac{1}{m} \cdot \|\tilde{x}_k - \tilde{x}_{k-1}\|_2^2 / (\tilde{x}_k - \tilde{x}_{k-1})^\top (g_k - g_{k-1})$
  end if
  $x_0 = \tilde{x}_k$
  for $t = 0, \cdots, m-1$ **do**
    Randomly pick $i_t \in \{1, \ldots, n\}$
    $x_{t+1} = x_t - \eta_k (\nabla f_{i_t}(x_t) - \nabla f_{i_t}(\tilde{x}_k) + g_k)$
  end for
  Option I: $\tilde{x}_{k+1} = x_m$
  Option II: $\tilde{x}_{k+1} = x_t$ for randomly chosen $t \in \{1, \ldots, m\}$
end for

5. Implementation

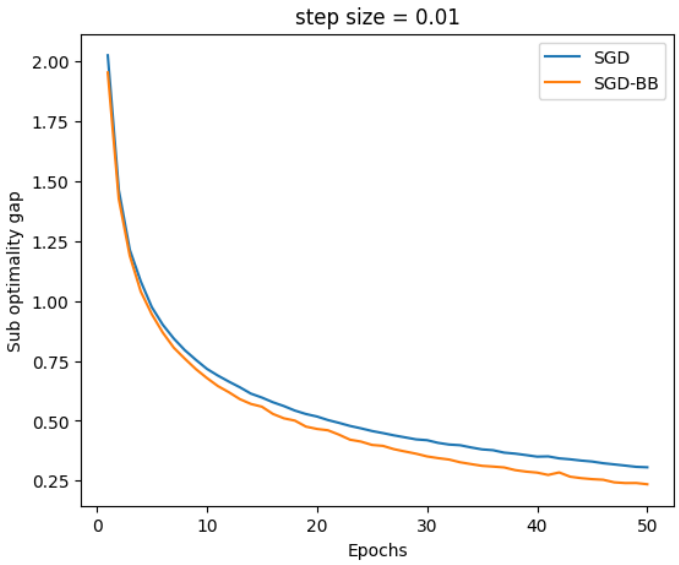
5.1. Data Set:

We took Fashion MNIST dataset to perform both SGD SVRG with BB Step Size.

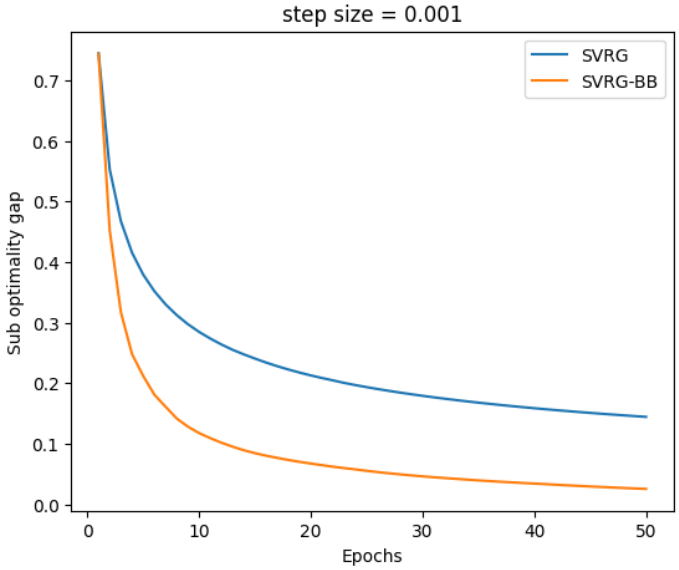
5.2. Results:

Intaillly i found f_opt by training the data set with Linear Regression using Sklearn module.Then I plot the graphs on basis of suboptimality gap between f_opt and f_pred for 50 epochs by using SGD SVRG.

5.2.1 SGD with BB :



5.2.2 SVRG with BB :



5.3. Observations :

Here iam using cross entropy for loss function.As per this paper we have to get suboptimality gaps as linear but here iam getting as sublinear but BB step size descent methods are converging faster than normal step size descent methods. This is due to may be the data set and intial step sizes. But BB provides better results than normal SGD SVRG.

5.4. Code :

This is link to our code Optml Project -2.

6. References :

1 . Barzilai-Borwein Step Size for Stochastic Gradient Descent Conghui Tan† Shiqian Ma† Yu-Hong Dai‡ Yuqiu Qian§ -2016

2 . J. Barzilai and J. M. Borwein. Two-point step size gradient methods. IMA Journal of Numerical Analysis, 8(1):141–148, 1988.