Optimization for Data Science

F. Rinaldi[1]

1 Università degli Studi di Padova — Dipartimento Matematica

Padova
2020

Outline

Optimization for Data Science

Block Coordinate Gradient Descent with Gauss-Southwell Rule

- We first analyze a block coordinate gradient with the Gauss-Southwell rule.
- We use a fixed stepsize $\alpha_k = 1/L$ in our analysis.

Scheme of the Algorithm

Algorithm 1 `Gauss-Southwell BCGD method`

Choose a point $x_1 \in \mathbb{R}^n$
For $k = 1, \ldots$
    If x_k satisfies some specific condition, then STOP
    Pick block i_k such that $i_k = \underset{j \in \{1,\ldots,b\}}{\text{Argmax}} \|\nabla_j f(x_k)\|$.
    Set
$$x_{k+1} = x_k - \frac{1}{L} U_{i_k} \nabla_{i_k} f(x_k)$$
End for

Assumption

Assumption 5 [Lipschitz Continuity]

- f has Lipschitz continuous gradient, with constant L;
- $f(., \mathbf{x}_{-i})$ has Lipschitz continuous gradient with constant L_i, that is

$$\|\nabla f(x + U_i h_i) - \nabla f(x)\| \leq L_i \|h_i\|, \text{ for all } h_i \in \mathbb{R}^{n_i} \text{ and } x \in \mathbb{R}^n.$$

- We also denote with $L_{max} = \max_i L_i$ and $L_{min} = \min_i L_i$.
- It is possible to see that

$$L_i \leq L \leq \sum_i L_i \leq b \cdot L_{max}, \forall\, i \in \{1, \ldots, b\}.$$

Convergence Results

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function satisfying Assumption 5. Gauss-Southwell BCGD method, satisfies:

$$f(x_{k+1}) - f(x^*) \leq \frac{2Lb\|x_1 - x^*\|^2}{k}.$$

If we further have that $f : \mathbb{R}^n \to \mathbb{R}$ is a σ-strongly convex function, then Gauss-Southwell BCGD method satisfies:

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\sigma}{bL}\right)^k (f(x_1) - f(x^*)).$$

- Rates similar to gradient method (b shows up in rates).
- Lipschitz assumption needs to be satisfied.

Proof

Taking into account reasoning seen for gradient method, we have

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L} \|\nabla_{i_k} f(x_k)\|^2.$$

Now, considering that $i_k = \underset{j \in \{1,\dots,b\}}{\text{Argmax}} \|\nabla_j f(x_k)\|$, we have

$$\|\nabla_{i_k} f(x_k)\|^2 \geq \frac{1}{b} \|\nabla f(x_k)\|^2 = \frac{1}{b} \sum_{i=1}^{b} \|\nabla_i f(x_k)\|^2.$$

Hence, we can write

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L} \|\nabla_{i_k} f(x_k)\|^2 \leq -\frac{1}{2Lb} \|\nabla f(x_k)\|^2.$$

Rest of the proof similar to the gradient descent method (the only difference is the term b that shows up here).

Comments

PROs

Good rates when proper conditions are met.

CONs

Block choice for updates requires:

- evaluating the whole gradient;
- searching for the best index.

- Very costly when tackling huge scale problems coming from specific data science applications.
- To practically implement Gauss-Southwell methods, some terms of those “greedy” scores may be cached and maintained at each iteration.

Sparse Optimization Problems

- Gauss-Southwell coordinate selection is very efficient for sparse optimization (i.e., optimization problems with sparse solutions).
- Most zero components in the solution are never selected and thus keep being zero throughout the iterations.
- Problem dimension effectively reduces to updated variables.
- The algorithm converges in very few iterations in practice.
- The saved iterations may over-weight the extra cost of ranking the coordinates.
- Smart update of gradient when problem has some structure!

Main Features

- Use the random sampling rule to choose the block at iteration k.
- Use again a fixed stepsize $\alpha_k = 1/L$ in theoretical analysis.

Expectation (Expected Value)

- The *expectation* or expected value of a random variable is a single number that tells you a lot about the behavior of the variable itself.
- Roughly speaking, the expectation is the average value of the random variable where each value is weighted according to its probability.

Expectation (Expected Value)

Definition (Expected Value of Continuous Variable)

Let X be a continuous random variable. The expected value of X is

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x \cdot P(x) dx.$$

Definition (Expected Value of Discrete Variable)

Let X be a discrete random variable. The expected value of X is

$$\mathbb{E}[X] = \sum_{x} x \cdot P(X = x).$$

Expectation (Expected Value)

Definition (Expected Value of Functions)

Let X be a discrete random variable, and let g be a function. The expected value of $g(X)$ is

$$\mathbb{E}[g(X)] = \sum_{x} g(x) \cdot P(X = x).$$

Algorithmic Scheme

Algorithm 2 `Randomized BCGD method`

Choose a point $x_1 \in \mathbb{R}^n$
For $k = 1, \ldots$
    If x_k satisfies some specific condition, then STOP
    Randomly pick $i_k \in \{1, \ldots, b\}$
    Set

$$x_{k+1} = x_k - \alpha_k U_{i_k} \nabla_{i_k} f(x_k)$$

End for

Details

- We consider a uniform distribution to randomly pick the block.
- Since i_k is a discrete random variable we have that

$$P(i_k = i) = \frac{1}{b}, \quad \forall\, i = 1, \dots, b.$$

- Keep in mind that the variables $i_1, \dots, i_k$ are all independent.

Convergence of Randomized BCGD Method

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function satisfying Assumption 5. Randomized BCGD method with $\alpha_k = \frac{1}{L}$, satisfies:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \frac{2LbR(x_1)^2}{k},$$

where $R(x_1)$ is

$$R(x_1) = \max\{\|x - x^*\|, f(x) \leq f(x_1), \ x \in \mathbb{R}^n\}.$$

If we further have that $f : \mathbb{R}^n \to \mathbb{R}$ is a σ-strongly convex function, then Randomized BCGD method satisfies:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \left(1 - \frac{\sigma}{bL}\right)^k (f(x_1) - f(x^*)).$$

Proof of Convergence

- Algorithm generates a random output $(x_{k+1}, f(x_{k+1}))$, which depends on the observed implementation of the random variable
$$\xi_k = \{i_1, i_2, \ldots, i_k\}.$$
- We show that the expected value $\mathbb{E}_{\xi_k}[f(x_{k+1})]$ converges to $f(x^*)$.

First of all, considering inequality from gradient method, we have

$$f(x_k - \frac{1}{L}U_i \nabla_i f(x_k)) - f(x_k) \leq -\frac{1}{2L}\|\nabla_i f(x_k)\|^2.$$

Taking the expected value with respect to i_k, and keeping in mind that $P(i_k = i) = \frac{1}{b}$, we get

$$\begin{aligned}
\mathbb{E}_{i_k}[f(x_{k+1})] - f(x_k) &= \sum_{i=1}^{b} \frac{1}{b}\left(f\left(x_k - \frac{1}{L}U_i \nabla_i f(x_k)\right) - f(x_k)\right) \qquad (1)\\
&\leq -\frac{1}{2L}\sum_{i=1}^{b}\frac{1}{b}\|\nabla_i f(x_k)\|^2 = -\frac{1}{2Lb}\|\nabla f(x_k)\|^2.
\end{aligned}$$

Proof of Convergence II

Considering first order convexity conditions, we can write

$$f(x_k) \leq f(x^*) + \nabla f(x_k)(x_k - x^*)$$

and thus we obtain, from Cauchy-Schwarz inequality the following:

$$f(x_k) - f(x^*) \leq \|\nabla f(x_k)\| \cdot \|x_k - x^*\|.$$

By using the fact that $f(x_{k+1}) \leq f(x_k)$, we get that

$$f(x_k) - f(x^*) \leq \|\nabla f(x_k)\| \cdot \|x_k - x^*\| \leq \|\nabla f(x_k)\| \cdot R(x_1).$$

that is

$$\|\nabla f(x_k)\| \quad \geq \quad \frac{f(x_k) - f(x^*)}{R(x_1)}.$$

Thus we obtain, by combining previous inequality with (1), the following:

$$\mathbb{E}_{i_k}[f(x_{k+1})] - f(x_k) \leq -\frac{1}{2Lb}\left(\frac{f(x_k) - f(x^*)}{R(x_1)}\right)^2. \qquad (2)$$

Proof of Convergence III

Now we use the definition of expectation to get

$$\mathbb{E}[f(x_{k+1})] = \mathbb{E}_{\xi_{k-1}}\left[\mathbb{E}_{i_k}\left[f(x_{k+1})\right]\right]$$

and the fact that

$$\mathbb{E}[f(x_k)] = \mathbb{E}_{\xi_{k-1}}[f(x_k)].$$

In order to prove first part of our result, we consider the expectation in ξ_{k-1} for both sides of the previous inequality:

$$\mathbb{E}_{i_k}[f(x_{k+1})] - f(x_k) \leq -\frac{1}{2Lb}\left(\frac{f(x_k) - f(x^*)}{R(x_1)}\right)^2.$$

Thus, using again basic properties of expectation, we get

$$\mathbb{E}[f(x_{k+1})] - \mathbb{E}[f(x_k)] \leq -\frac{1}{2Lb}\frac{\mathbb{E}\left[\left(f(x_k) - f(x^*)\right)^2\right]}{R(x_1)^2} \leq -\frac{1}{2Lb}\frac{\left(\mathbb{E}[f(x_k) - f(x^*)]\right)^2}{R(x_1)^2}.$$

Proof of Convergence IV

Now, consider inequality

$$\mathbb{E}[f(x_{k+1})] - \mathbb{E}[f(x_k)] \leq -\frac{1}{2Lb}\frac{\left(\mathbb{E}[f(x_k) - f(x^*)]\right)^2}{R(x_1)^2}.$$

We finally call $r_k = \mathbb{E}[f(x_k)] - f(x^*)$ and $\gamma = \frac{1}{2LbR(x_1)^2}$. Hence we have

$$r_{k+1} - r_k \leq -\gamma r_k^2.$$

The rest of the proof follows from analysis of gradient descent method.

Improving the Rate

- There exist different strategies to improve the rate of the randomized BCGD algorithm.
- **First idea**: use larger stepsizes (replace $\alpha_k = 1/L$ with $\alpha_k = 1/L_{i_k}$).
- **Second idea**: use non-uniform sampling.
- Nesterov's idea: use

$$P(i_k = i) = \frac{L_i}{\sum_{i=1}^{b} L_i},$$

i.e. we choose block with larger Lipschitz constant more frequently.

Convergence Rate with Non-uniform Sampling

Proposition

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function satisfying Assumption 5. Randomized BCGD method with $\alpha_k = 1/L_i$ and non-uniform sampling distribution satisfies:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \frac{2\sum_i L_i R(x_1)^2}{k},$$

where $R(x_1)$ is

$$R(x_1) = \max\{\|x - x^*\|, \, f(x) \leq f(x_1), \; x \in \mathbb{R}^n\}.$$

If we further have that $f : \mathbb{R}^n \to \mathbb{R}$ is a σ-strongly convex function, then Randomized BCGD method with $\alpha_k = 1/L_i$ and non-uniform sampling distribution satisfies:

$$\mathbb{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \left(1 - \frac{\sigma}{\sum_i L_i}\right)^k (f(x_1) - f(x^*)).$$

Convergence Rate with Non-uniform Sampling II

Remark

When $L_i = L/b,\ i = 1, \dots, b$, modified version of the randomized BCGD method achieves same complexity result as full gradient descent algorithm

$\dots$ but the iteration cost is much cheaper!

Comments

PROs

- Computation of a partial derivative is much cheaper and less memory demanding than computing the whole gradient.
- Randomized BCGD is well suited when memory is limited.
- Randomization improves the convergence rate of BCGD in expectation.

Comments

CONs

- Randomized algorithms have to sample from probability distributions ($\mathcal{O}(n)$ operations) at each iteration.
- For huge-scale problems this complexity can be prohibitive (alternative strategies with complexity $\mathcal{O}(\ln n)$).
- Randomized BCGD variants might have bigger iteration complexities than cyclic BCGD methods.
- Results in practice may vary depending on the runs.
- Cache misses are more likely (requiring extra time to move data from slower to faster memory in the memory hierarchy).

Cyclic Rule

- We consider the cyclic rule to make the updates at iteration k.
- We use again a fixed stepsize $\alpha_i = 1/L$ in our analysis.

Cyclic BCGD Method

Algorithm 3 `Cyclic BCGD method`

Choose a point $x_1 \in \mathbb{R}^n$
For $k = 1, \ldots$
    If x_k satisfies some specific condition, then STOP
    Set $y_0 = x_k$
    For $i = 1, \ldots, b$, set

$$y_i = y_{i-1} - \alpha_i U_i \nabla_i f(y_{i-1})$$

    Set $x_{k+1} = y_b$
End for

Convergence of Cyclic BCGD

Convergence of BCGD with cyclic rule

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function satisfying Assumption 5. Cyclic BCGD method with $\alpha_k = \frac{1}{L}$, satisfies:

$$f(x_{k+1}) - f(x^*) \leq \frac{4L(b+1)R(x_1)^2}{k},$$

where $R(x_1)$ is

$$R(x_1) = \max\{\|x - x^*\|, f(x) \leq f(x_1), \; x \in \mathbb{R}^n\}.$$

If we further have that $f : \mathbb{R}^n \to \mathbb{R}$ is a σ-strongly convex function, then Cyclic BCGD method satisfies:

$$f(x_{k+1}) - f(x^*) \leq \left(1 - \frac{\sigma}{2(b+1)L}\right)^k (f(x_1) - f(x^*)).$$

Proof

We can write

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{4L(b+1)} \|\nabla f(x_k)\|^2. \tag{3}$$

Considering first order convexity conditions, we can write

$$f(x_k) \leq f(x^*) + \nabla f(x_k)(x_k - x^*)$$

and thus we obtain, from Cauchy-Schwarz inequality the following:

$$f(x_k) - f(x^*) \leq \|\nabla f(x_k)\| \cdot \|x_k - x^*\|.$$

By using the fact that $f(x_{k+1}) \leq f(x_k)$, we get that

$$f(x_k) - f(x^*) \leq \|\nabla f(x_k)\| \cdot \|x_k - x^*\| \leq \|\nabla f(x_k)\| \cdot R(x_1).$$

that is

$$\|\nabla f(x_k)\| \quad \geq \quad \frac{f(x_k) - f(x^*)}{R(x_1)}.$$

Proof II

Thus we obtain, by combining previous inequality with (3), the following:

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{4L(b+1)} \left(\frac{f(x_k) - f(x^*)}{R(x_1)} \right)^2 . \tag{4}$$

We now call $r_k = f(x_k) - f(x^*)$ and $\gamma = \frac{1}{4L(b+1)R(x_1)^2}$. Hence we have

$$r_{k+1} - r_k \leq -\gamma r_k^2 .$$

The rest of the proof directly follows from the analysis we carried out for the gradient descent method.

Comments

- The cyclic BCGD method is a deterministic algorithm.
- Its iteration cost is $\mathcal{O}(b)$ times larger than the randomized BCGD method.
- Cyclic variants are most intuitive and easily implemented.
- BCGD with the deterministic cyclic rule poorer performance than that with randomized cyclic one.
- Rates improved by using a better stepsize, i.e., $\alpha_i = 1/L_i$. LCG case we get

$$f(x_{k+1}) - f(x^*) \leq \frac{4L_{max}(bL^2/L_{min}^2 + 1)R(x_1)^2}{k}.$$

PROs and CONs

PROs

- The cyclic BCGD method is a deterministic algorithm.
- Cyclic variants are most intuitive and easily implemented.

CONs

- Iteration cost is $\mathcal{O}(b)$ times larger than randomized BCGD.
- Rates are worse than the other BCGD methods