CSC 576: Coordinate Descent Algorithms

Ji Liu
Department of Computer Sciences, University of Rochester

November 19, 2015

1 Coordinate descent and block coordinate descent

The coordinate descent (CD) algorithm is also a popular algorithm in optimization. For any optimization problems $\min_{x\in\mathbb{R}^n} \ f(x)$, we can split the vector x into n coordinates and rewrite the optimization problem in below

$$\min_{x_1,x_2,\cdot,x_n} \quad f(x_1, x_2, \cdots, x_n).$$

(Note that $x_1, \cdots, x_n$ denote the coordinates of x, which is different from our previous note). The idea of coordinate descent is to iteratively optimize each coordinate x_i while fix the rest. The algorithm can be summarized in 1: A slightly more general version of CD is the block version: block

Algorithm 1 CD

Require: x_0, K
Ensure: x_K;
  for $k = 1 : K$ **do**
    for $i = 1 : n$ **do**

$$x_i = arg\min_{y} \quad f(x_1, \cdots, x_{i-1}, y, x_i, \cdots, x_n) \tag{1}$$

    end for
  end for

coordinate descent (BCD). Following the same spirit in CD, BCD optimizes a block of coordinates (instead of a single) while fixing the rest. One example of using BCD is the matrix factorization problem covered in our early note:

$$\min_{U,V} \quad \frac{1}{2}\|M - UV^T\|_F^2.$$

We optimize U and V iteratively by

$$U = M(V^+)^T \quad V = M^T(U^+)^T.$$

Since there is not any key difference between BCD and CD. When we say CD, it covers both cases. Many well known algorithms such as K-means and EM can be considered as a special case of CD.

We can see that every single iterate decreases the objective function. Therefore, it is easy to say that the CD algorithm guarantees the convergence. However, it does not guarantee to converge to the global optimum even for convex problems. For example, consider the following problem:

$$\min_{x,y} \quad \frac{1}{2}x^2 + |x - y|.$$

It is easy to see that the optimal solution is $(0, 0)$. However, if (x, y) are initialized to $(1, 1)$, we have

$$\begin{aligned} y =& 1 = arg\min_{y} \ \frac{1}{2}1^2 + |1 - y| \quad \text{fix } x = 1 \\ x =& 1 = arg\min_{x} \ \frac{1}{2}x^2 + |x - 1| \quad \text{fix } y = 1. \end{aligned}$$

To guarantee that CD converges to the optimal solution, we need some additional conditions.

- $f(x)$ is a convex smooth function;
- $f(x)$ consists of a convex smooth function $F(x)$ plus a convex closed "separable" function $G(x)$. $G(x)$ is separable means that $G(x) = \sum_{i=1}^{n} G_i(x_i)$.

2 Gauss-Seidel CD

Sometimes, it is still difficult to minimize a single variable (or a single block of variables) in CD. However, that is not a big deal. The sprite of CD is nothing but updating a single variable to decrease the objective value every time. Therefore, it is not necessary to optimize each single variable to the best. To ensure to decrease the objective value, a simple coordinate gradient step is enough. The Gauss-Seidel CD updates a coordinate by

$$x_i = x_i - \gamma \nabla_i f(x)$$

where $\nabla_i f(x)$ denotes the ith coordinate of the gradient $\nabla_i f(x)$. The steplength γ should be chosen properly such that the new x_i decreases the objective function value. A safe way to choose γ is

$$\gamma = \frac{1}{\max_x |\nabla^2_{ii} f(x)|} \tag{2}$$

To adaptively choose the step length, one can use the line search scheme we used in gradient descent method. If we use (2) to replace (1), we obtain the Gauss-Seidel CD. Now, when we say CD, it covers Gauss-Seidel CD as well.

3 Stochastic CD

The standard CD algorithm updates coordinates (or blocks of coordinates) in the cyclic manner order. It may be extremely slow in practice. The stochastic version of CD updates coordinates in a randomize manner in Algorithm 2.

Algorithm 2 Stochastic CD

Require: x_0, K
Ensure: x_K;
  for $k = 1 : K$ **do**
    Randomly select a coordinate i from $\{1, 2, \cdots, n\}$
    Update x_i by (1) or (2)
  end for

4 EM algorithm: An explanation using CD

The key principle of CD is to optimize a single variable or (a single group of variable) at each iteration while the remaining variables are fixed. We can use it to explain the EM algorithm. The general framework of EM is trying to maximize the likelihood $\mathbb{P}_\theta(X)$, which can be reformulated in the following:

$$\begin{aligned}
&\max_\theta \ \mathbb{P}_\theta(X) \\
\Leftrightarrow &\max_\theta \ \log \mathbb{P}_\theta(X) \\
\Leftrightarrow &\max_\theta \ \log \mathbb{P}_\theta(X) + \max_{q(\cdot)} \sum_Z q(Z) \log \frac{\mathbb{P}_\theta(Z|X)}{q(Z)} \\
\Leftrightarrow &\max_{\theta, q(\cdot)} \ \sum_Z q(Z) \log \mathbb{P}_\theta(X) + \sum_Z q(Z) \log \frac{\mathbb{P}_\theta(Z|X)}{q(Z)} \\
\Leftrightarrow &\max_{\theta, q(\cdot)} \ \sum_Z q(Z) \left(\log \mathbb{P}_\theta(X) + \log \frac{\mathbb{P}_\theta(Z|X)}{q(Z)} \right) \\
\Leftrightarrow &\max_{\theta, q(\cdot)} \ \sum_Z q(Z) \left(\log \mathbb{P}_\theta(X) + \log \frac{\mathbb{P}_\theta(Z|X)}{q(Z)} \right) \\
\Leftrightarrow &\max_{\theta, q(\cdot)} \ \sum_Z q(Z) \log \frac{\mathbb{P}_\theta(Z, X)}{q(Z)}
\end{aligned}$$

where $q(\cdot)$ is a density function, and the third line uses the fact of KL-divergence

$$\sum_Z q(Z) \log \frac{p(Z)}{q(Z)} = -KL(p||q) \leq 0$$

and the equality holds when $p(Z) = q(Z)$. The EM algorithm is nothing but to optimize the objective function by alternatively optimizing θ and $q(\cdot)$

- optimize $q(\cdot)$ by
$$q(Z) = \mathbb{P}_\theta(Z|X) = \operatorname*{argmax}_{q(\cdot)} \ \sum_Z q(Z) \log \frac{\mathbb{P}_\theta(Z, X)}{q(Z)};$$
- optimize θ by
$$\theta = \operatorname*{argmax}_\theta \ \sum_Z q(Z) \log \frac{\mathbb{P}_\theta(Z, X)}{q(Z)}.$$