CD

Coordinate Descent Algorithms

Liu Jingchang

March 23, 2017

University of Science and Technology of China

Table of contents

Algorithms

Motivation

Stochastic Gradient Descent

$$n \gg p \quad f(x) = \sum_{i=1}^{n} f_i(x) \tag{1}$$

Coordinate Descent

p may big, such as in computational biology and healthy care problem.

Formulation

Unconstrained minimization

$$\min_x f(x) \tag{2}$$

Structured formulation

$$\min_x h(x) := f(x) + \lambda\Omega(x) \tag{3}$$

- f: smooth
- Ω: regularization function, may be nonsmooth, often convex, usually assumed to be separable or block-separable.

$$\Omega(x) = \sum_{i=1}^{n} \Omega_i(x_i) \tag{4}$$

- $\Omega(x) = \|x\|_1$, $\Omega_i(x_i) = \|x_i\|_1$
- Ω: box constraints, $\Omega_i(x_i) = I_{[l_i, u_i]}(x_i)$

Algorithm1

Algorithm 1 Coordinate Descent for unconstrained case

1: Set $k \leftarrow 0$ and choose $x^0 \in \mathbb{R}^n$
2: **repeat**
3: Choose index $i_k \in \{1, 2, ..., n\}$
4: $x^{k+1} \leftarrow x^k - \alpha_k \nabla_{i_k} f\left(x^k\right) e_{i_k}$
{equal to $x_{i_k}^{k+1} \leftarrow x_{i_k}^k - \alpha_k \nabla_{i_k} f\left(x^k\right)$}
5: $k \leftarrow k + 1$
6: **until** termination test satisfied

Proximal

Formulation

$$\min_{x}\ h(x) := f(x) + \lambda\Omega(x) \tag{5}$$

f: smooth, Ω: may be nonsmooth.

Iteration

$$\begin{aligned} x^{k+1} &= prox_{\Omega}^{\eta}\left(x^{k} - \eta\nabla f\left(x^{k}\right)\right) & (6) \\ &= \operatorname*{argmin}_{y}\left\{\Omega(y) + \frac{1}{2\eta}\left\|y - \left(x^{k} - \eta\nabla f\left(x^{k}\right)\right)\right\|_{2}^{2}\right\} & (7) \end{aligned}$$

Indicate function

$$
\begin{aligned}
x^{k+1} =& prox_{\Omega}^{\eta}\left(x^k - \eta\nabla f\left(x^k\right)\right) & (8)\\
=& \underset{y}{\operatorname{argmin}}\left\{\Omega\left(y\right) + \frac{1}{2\eta}\left\|y - \left(x^k - \eta\nabla f\left(x^k\right)\right)\right\|_2^2\right\} & (9)\\
=& \underset{y\in\mathbb{C}}{\operatorname{argmin}}\left\{\left\|y - \left(x^k - \eta\nabla f\left(x^k\right)\right)\right\|_2^2\right\} & (10)\\
=& \operatorname{Proj}_{\mathbb{C}}\left(x^k - \eta\nabla f\left(x^k\right)\right) & (11)
\end{aligned}
$$

Other

- l_1 norm: soft-thresholding
- l_0 norm: hard-thresholding

Figure

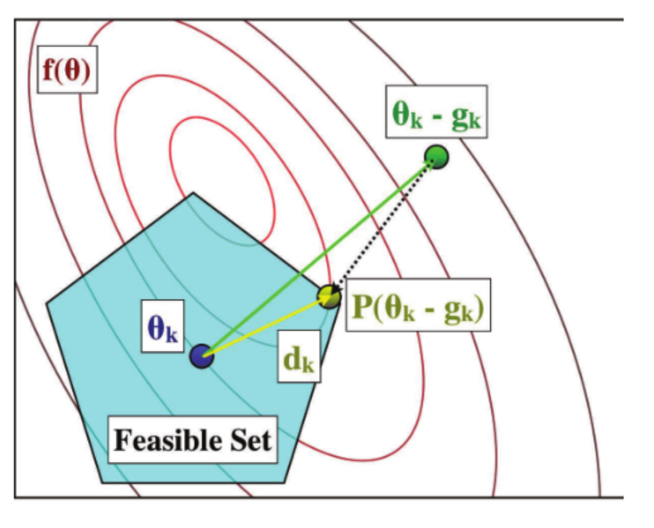


Figure 1: Illustration of projected gradient descent

Algorithm

Algorithm 2 Coordinate Descent for regularization case

1: Set $k \leftarrow 0$ and choose $x^0 \in \mathbb{R}^n$
2: **repeat**
3:     Choose index $i_k \in \{1, 2, ..., n\}$
4:     $z_{i_k}^{k+1} \leftarrow prox_{\Omega_{i_k}}^{\eta}\left(x_{i_k}^k - \eta \nabla_{i_k} f\left(x^k\right)\right)$
5:     $x^{k+1} \leftarrow x^k + \left(z_{i_k}^k - x_{i_k}^k\right) e_{i_k}$
    {4 and 5 is equal to $x_{i_k}^{k+1} \leftarrow prox_{\Omega_{i_k}}^{\eta}\left(x_{i_k}^k - \eta \nabla_{i_k} f\left(x^k\right)\right)$}
6:     $k \leftarrow k + 1$
7: **until** termination test satisfied

Property

$$prox_{\Omega}^{\eta}\left(x^k - \eta \nabla f\left(x^k\right)\right) = \left(prox_{\Omega_1}^{\eta}\left(x_1^k - \eta \nabla_1 f\left(x^k\right)\right), prox_{\Omega_2}^{\eta}\left(x_2^k - \eta \nabla_2 f\left(x^k\right)\right), \cdots, prox_{\Omega_n}^{\eta}\left(x_n^k - \eta \nabla_n f\left(x^k\right)\right)\right) \tag{12}$$

Analysis

Gradient Descent

Full Gradient Descent

$$\min_x f(x) \qquad x^{k+1} \leftarrow x^k - \eta_k \nabla f\left(x^k\right)$$

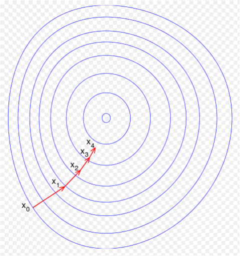
Stochastic Gradient Descent

$$\min_x \sum_{i=1}^{n} f_i(x) \qquad x^{k+1} \leftarrow x^k - \eta_k \nabla f_{i_k}\left(x^k\right)$$

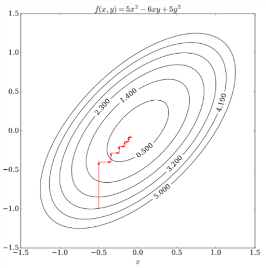
Stochastic Coordinate Descent

$$\min_x f(x) \qquad x^{k+1} \leftarrow x^k - \eta_k \nabla_{i_k} f\left(x^k\right)$$

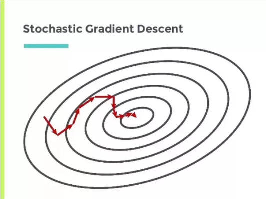
Gradient Descent



SCD



SGD



Assumption

Strong convexity

$$\exists\ \delta > 0,\ sit. f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\delta}{2} \|y - x\|_2^2, \quad \text{for all } x, y \tag{13}$$

Component Lipschitz Continuous

$$\|\nabla_i f(x + te_i) - \nabla_i f(x)\| \leq L_i \|t\| \tag{14}$$

Standard Lipschitz Continuous

$$\|\nabla f(x + d) - \nabla f(x)\| \leq L \|d\| \tag{15}$$

Assumption

f: Convex, Lipschitz continuous gradient, give $x^0, \exists\ R_0$, sit.

$$\max_{x^* \in \mathcal{S}} \max_{x} \left\{ \|x - x^*\| : f(x) \leq f\left(x^0\right) \right\} \leq R_0 \tag{16}$$

Convergence rate of unconstrained case

Th

Under above assumption. Suppose that $\alpha_k = \frac{1}{L_{max}}$, then

$$E\left(f\left(x^k\right)\right) - f^* \leq \frac{2nL_{max}R_0^2}{k} \tag{17}$$

When δ strongly-convex,

$$E\left(f\left(x^k\right)\right) - f^* \leq \left(1 - \frac{\delta}{nL_{max}}\right)^k \left(f\left(x^0\right) - f^*\right) \tag{18}$$

Proof

$$
\begin{aligned}
f\left(x^{k+1}\right) =& f\left(x^k - \alpha_k \nabla_{i_k} f\left(x^k\right) e_{i_k}\right) \\
\leq& f\left(x^k\right) - \alpha_k \left[\nabla_{i_k} f\left(x^k\right)\right]^2 + \frac{1}{2}\alpha_k^2 L_{i_k} \left[\nabla_{i_k} f\left(x^k\right)\right]^2 \\
\leq& f\left(x^k\right) - \alpha_k \left(1 - \frac{L_{max}}{2}\alpha_k\right) \left[\nabla_{i_k} f\left(x^k\right)\right]^2 \\
=& f\left(x^k\right) - \frac{1}{2L_{max}} \left[\nabla_{i_k} f\left(x^k\right)\right]^2 \qquad (19)
\end{aligned}
$$

Taking the expectation of both sides over the random index i_k,

$$
\begin{aligned}
E_{i_k} f\left(x^{k+1}\right) \leq& f\left(x^k\right) - \frac{1}{2L_{max}} \frac{1}{n} \sum_{i=1}^{m} \left[\nabla_i f\left(x^k\right)\right]^2 \\
\leq& f\left(x^k\right) - \frac{1}{2nL_{max}} \left[\nabla f\left(x^k\right)\right]^2 \qquad (20)
\end{aligned}
$$

Proof

Take expectation with respect to $i_1, i_2, \cdots, i_{k-1}$,

$$E\left(f\left(x^{k+1}\right)\right) \leq E\left(f\left(x^{k}\right)\right) - \frac{1}{2nL_{max}} E\left[\nabla f\left(x^{k}\right)\right]^{2} \tag{21}$$

By the convexity of f

$$f\left(x^{k}\right) - f^{*} \leq \nabla f\left(x^{k}\right)^{T}\left(x^{k} - x^{*}\right) \leq \left\|\nabla f\left(x^{k}\right)\right\| \left\|x^{k} - x^{*}\right\| \leq R_0 \left\|\nabla f\left(x^{k}\right)\right\| \tag{22}$$

Combine 21 and 22, we can get 17.

Proof for strong convex

Strong convex

$$\exists\delta > 0, sit. f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\delta}{2} \|y - x\|_2^2, \quad \text{for all } x, y$$

Taking the minimum of both sides with respect to y, and setting $x = x^k$

$$f^* \geq f\left(x^k\right) - \frac{1}{2\alpha} \left\|\nabla f\left(x^k\right)\right\|^2 \tag{23}$$

Combine with 18, we obtain

$$E_{i_k} f\left(x^{k+1}\right) - f^* \leq f\left(x^k\right) - \frac{\delta}{nL_{max}} f\left(x^k\right) \tag{24}$$

Take expectation with respect to $i_1, i_2, \cdots, i_{k-1}$,

$$E\left(f\left(x^{k+1}\right)\right) - f^* \leq E\left(f\left(x^k\right)\right) - f^* - \frac{\delta}{nL_{max}} \left(E\left(f\left(x^k\right)\right) - f^*\right) \tag{25}$$

Assumption

- f : Lipschitz continuous gradient, Strongly convex with $\delta > 0$
- $\Omega_i, i = 1, 2, \cdots, n$: convex

Th

Under above assumption, $\alpha_k = \frac{1}{L_{max}}$, Then

$$E\left(h\left(x^k\right)\right) - h^* \leq \left(1 - \frac{\delta}{nL_{max}}\right)^k \left(h\left(x^0\right) - h^*\right) \tag{26}$$

Conclusion

Conclusion

- Coordinate Descent is an approximation to Full Gradient Descent.
- Proximal methods is designed for regularized optimization.
- CD an get $O\left(\frac{1}{n}\right)$ convergence; And linear convergence when strongly convex.

Q & A