Coordinate Descent Algorithms (SCD): Gauss-Southwell Rule

Jingchang Liu

December 16, 2018

University of Science and Technology of China

Table of Contents

Introductions

Gauss-Southwell Rule

Conclusion

Q & A

Introductions

Motivation

Stochastic Gradient Descent

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

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

Gradient Descent

Full Gradient Descent

$$\min_x f(x) \qquad x^{k+1} \leftarrow x^k - \gamma_k \nabla f(x^k).$$

Stochastic Gradient Descent

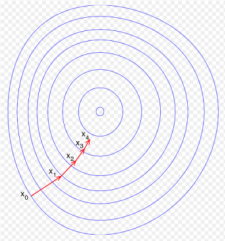
$$\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x) \qquad x^{k+1} \leftarrow x^k - \gamma_k \nabla f_{i_k}(x^k).$$

Stochastic Coordinate Descent

$$\min_x f(x) \qquad x^{k+1} \leftarrow x^k - \gamma_k \nabla_{i_k} f(x^k).$$

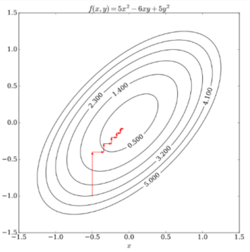
Figure

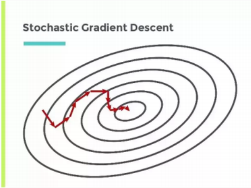
Gradient Descent



SCD

SGD





SDCA

Formulation

$$\min_{w \in \mathbb{R}^d} P\left(w\right)$$

$$P\left(w\right) := \frac{1}{n} \sum_{i=1}^{n} \phi_i \left(w^T x_i\right) + \frac{\lambda}{2} \left\|w\right\|^2$$

Dual Problem

$$\max_{\alpha} D\left(\alpha\right)$$

$$D\left(\alpha\right) = \frac{1}{n} \sum_{i=1}^{n} -\phi_i^* \left(-\alpha_i\right) - \frac{\lambda}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^{n} \alpha_i x_i \right\|^2$$

Conjugate function: $\phi_i^* \left(u\right) = \max_z \left(zu - \phi_i \left(z\right)\right)$

Derivation

$P(w) = \frac{1}{n}\sum\limits_{i=1}^{n} \phi_i\left(w^T x_i\right) + \frac{\lambda}{2}\|w\|^2$ equals to

$$
\begin{aligned}
P(y, z) &= \frac{1}{n}\sum_{i=1}^{n} \phi_i(z_i) + \frac{\lambda}{2}\|y\|^2 \\
s.t. &\quad y^T x_i = z_i, i = 1, 2, \cdots, n
\end{aligned}
$$

$$
L(y, z, \alpha) = P(y, z) + \frac{1}{n}\sum_{i=1}^{n} \alpha_i\left(y^T x_i - z_i\right)
$$

$$
\begin{aligned}
D(\alpha) &= \inf_{y,z} L(y, z, \alpha) \\
&= \frac{1}{n}\sum_{i=1}^{n} \inf_{z_i}\left\{\phi_i(z_i) - \alpha_i z_i\right\} + \inf_{y}\left\{\frac{\lambda}{2}\|y\|^2 + \frac{1}{n}\sum_{i=1}^{n} \alpha_i y^T x_i\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} -\phi_i^*(-\alpha_i) - \frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^{n} \alpha_i x_i\right\|^2.
\end{aligned}
$$

Algorithms

Let $w^{(0)} = w(\alpha^{(0)})$
Iterate: for $t = 1, 2, \dots, T$:
Randomly pick i
Find $\Delta\alpha_i$ to maximize $-\phi_i^*(-(\alpha_i^{(t-1)} + \Delta\alpha_i)) - \frac{\lambda n}{2}\|w^{(t-1)} + (\lambda n)^{-1}\Delta\alpha_i x_i\|^2$
$\alpha^{(t)} \leftarrow \alpha^{(t-1)} + \Delta\alpha_i e_i$
$w^{(t)} \leftarrow w^{(t-1)} + (\lambda n)^{-1}\Delta\alpha_i x_i$
Output (Averaging option):
Let $\bar{\alpha} = \frac{1}{T-T_0}\sum_{i=T_0+1}^{T} \alpha^{(t-1)}$
Let $\bar{w} = w(\bar{\alpha}) = \frac{1}{T-T_0}\sum_{i=T_0+1}^{T} w^{(t-1)}$
return $\bar{w}$

Figure 2: Procedure SDCA

Assumptions

Strong convexity

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

Component Lipschitz Continuous

$$\|\nabla_i f(x + te_i) - \nabla_i f(x)\| \leq L_i \|t\|.$$

Standard Lipschitz Continuous

$$\|\nabla f(x + d) - \nabla f(x)\| \leq L_i \|d\|.$$

Gauss-Southwell Rule

Block selected rules

Updates

$$x^{k+1} \leftarrow x^k - \frac{1}{L}\nabla_{i_k} f(x^k) e_{i_k}.$$

Random rule

$$i_k \text{ is selected randomly.}$$

GS

$$i_k = \arg\max_i |\nabla_i f(x^k)|.$$

Analysing

Lipschitz

$$
\begin{aligned}
& f(x^{k+1}) \\
\leq \quad & f(x^k) + \nabla_{i_k} f(x^k)(x^{k+1} - x^k)_{i_k} + \frac{L}{2}(x^{k+1} - x^k)_{i_k}^2 \\
= \quad & f(x^k) - \frac{1}{L}(\nabla_{i_k} f(x^k))^2 + \frac{L}{2}\left[\frac{1}{L}\nabla_{i_k} f(x^k)\right]^2 \\
= \quad & f(x^k) - \frac{1}{2L}\left[\nabla_{i_k} f(x^k)\right]^2.
\end{aligned}
$$

Strongly convex

$$
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|,
$$

which implies

$$
f(x^*) \geq f(x^k) - \frac{1}{2\mu}\|\nabla f(x^k)\|^2.
$$

Analysing for randomized selection

$$
\begin{aligned}
\mathbb{E}[f(x^{k+1})] &\leq \mathbb{E}\left[f(x^k) - \frac{1}{2L}(\nabla_{i_k} f(x^k))^2\right] \\
&= f(x^k) - \frac{1}{2L}\sum_{i=1}^{n}\frac{1}{n}(\nabla f(x^k))^2 \\
&= f(x^k) - \frac{1}{2Ln}\|\nabla f(x^k)\|^2.
\end{aligned}
$$

Using $f(x^*) \geq f(x^k) - \frac{1}{2\mu}\|\nabla f(x^k)\|^2$**:**

$$
\mathbb{E}[f(x^{k+1})] - f(x^*) \leq \left(1 - \frac{\mu}{Ln}\right)[f(x^k) - f(x^*)].
$$

Analysing for GS

$$(\nabla_{i_k} f(x^k))^2 = \|\nabla f(x^k)\|_\infty^2 \geq \frac{1}{n}\|\nabla f(x^k)\|^2.$$

Applying to $f(x^{k+1}) \leq f(x^k) - \frac{1}{2Ln}\|\nabla f(x^k)\|^2$**:**

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2Ln}\|\nabla f(x^k)\|^2.$$

Together with $f(x^*) \geq f(x^k) - \frac{1}{2\mu}\|\nabla f(x^k)\|^2$**:**

$$f(x^{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{Ln}\right)[f(x^k) - f(x^*)].$$

Almost the same convergence rate as that in randomized selection.

New analysis

New definition of Strong convexity

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_1}{2} \|y - x\|_1^2.$$

Minimizing both sides

$$\begin{aligned} f(x^*) &\geq f(x) - \sup_y \{ \langle -\nabla f(x), y - x \rangle - \frac{\mu_1}{2} \|y - x\|_1^2 \} \\ &= f(x) - \left(\frac{\mu_1}{2} \| \cdot \|_1^2 \right)^* (-\nabla f(x)) \\ &= f(x) - \frac{1}{2\mu_1} \|\nabla f(x)\|_\infty^2, \end{aligned}$$

which makes use of the convex conjugate $(\frac{\mu_1}{2} \| \cdot \|_1^2)^* = \frac{1}{2\mu_1} \| \| _\infty^*$.

Applying $(\nabla_{i_k} f(x^k))^2 = \|\nabla f(x^k)\|_\infty^2$:

$$f(x^{k+1}) - f(x^*) \leq \left(1 - \frac{\mu_1}{L} \right) [f(x^k) - f(x^*)].$$

Relationship between μ_1 and μ

Relationship between the 2-norm and the 1-norm

$$\|x\|_1 \geq \|x\| \geq \frac{1}{\sqrt{n}}\|x\|_1.$$

f is μ-strongly convex in the 2-norm

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \\ &\geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2n}\|y - x\|_1^2, \end{aligned}$$

implying that f is at least $\frac{\mu}{n}$-strongly convex in the 1-norm.

f is μ_1-strongly convex in the 1-norm

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu_1}{2}\|y - x\|_1^2 \\ &\geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu_1}{2n}\|y - x\|^2, \end{aligned}$$

implying that f is at least $\frac{\mu_1}{n}$-strongly convex in the 2-norm.

Conclusion of GS

Relationship between μ_1 and μ

$$\frac{\mu}{n} \leq \mu_1 \leq \mu.$$

Conclusions

- At one extreme the GS rule obtains the same rate as uniform selection ($\mu_1 \approx \mu/n$).
- GS may be faster than uniform selection by a factor of n ($\mu_1 \approx \mu$).

Lipschitz Sampling

Explanations

- L_i: how smooth the coordinate is.
- $p_i = L_i / \sum_{j=1}^n L_j$.

Convergence rate

$$f(x^{k+1}) - f(x^*) \leq \left(1 - \frac{\mu}{n\bar{L}}\right)[f(x^k) - f(x^*)],$$

where $\bar{L} = \frac{1}{n}\sum_{j=1}^n L_j$.

Remark

- This rate is faster than that for uniform sampling if any L_i differ.
- Is other distributions can make the rate faster? Like $p_i = \sqrt{L_i} / \sum_{j=1}^n \sqrt{L_j}$

Experiments

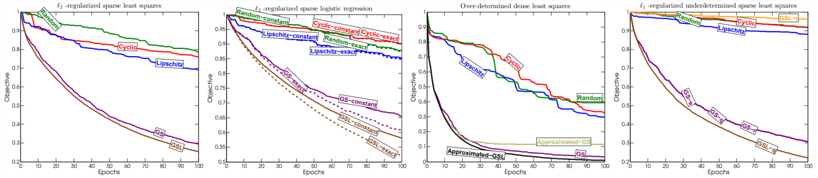


Figure 3: Comparison of coordinate selection rules.

Conclusion

Conclusions

- SCD can play its roles in some cases, like SDCA.
- GS is faster than uniform selection in almost all cases.
- More practical methods are needed to apply GS roles.

Q & A