Note 8.12

$\|\cdot\| := \|\cdot\|_2$ norm

1 Stochastic LM Algorithm

Consider the following least square problem

$$\min_{x \in R^d} f(x) := \frac{1}{2}\|r(x)\|^2 = \frac{1}{2}\sum_{i=1}^{M} |r_i(x)|^2 \tag{1}$$

where r_i are continuously differentiable, $i = 1, \cdots, M$. Build a set $\mathcal{F}_b := \{F_l \mid F_l \subseteq$ the power set of $\{1, \cdots, d\}$, and $|F_l| = b\}$, uniformly random choose $F \in \mathcal{F}_b$

$$\min f(x_F) := \frac{1}{2}\|r(x_F)\|^2 \tag{2}$$

Construct a quadratic function as follows

$$m_k(x_F^k + h) := f(x_F^k) + g^T(x_F^k)\,h + \frac{1}{2}h^T H(x_F^k)\,h + \frac{1}{2}\mu^k\|h\|^2 \tag{3}$$

where μ_k is the regularized parameter and $g(x_F^k) = \frac{1}{2}\sum_{i=1}^{M} \nabla |r_i(x_F)|^2 = \sum_{i=1}^{M} \nabla r_i(x_F)\, r_i(x_F)$, $H(x_F^k) = \sum_{i=1}^{M} \nabla r_i(x_F) \nabla\, r_i(x_F)^T$. the LM step is obtained by solving the subproblem: $\operatorname{argmin}_h m_k(x_F^k + h)$

i.e $h_F^k \leftarrow (H(x_F^k) + \mu^k I)h = -g(x_F^k)$. the ratio ρ is defined as

$$\rho_k = \frac{f(x_F^k) - f(x_F^k + h_F^k)}{m_k(x_F^k) - m_k(x_F^k + h_F^k)} \tag{4}$$

Algorithm 1.1 Stochastic LM algorithm

Step 0 Random choose an parameter set F and $\mu > 0$, initial damping parameter $\mu^0 = \mu\|r(x_F^0)\|^2$, constants $\gamma > 1$, $\mu_{\min}$ and $\eta_1, \eta_2 > 0$,set $k = 0$

Step 1 if a stopping criteria is satisfied, go to Step 0 or stop; otherwise, go to Step 2.

Step 2 obtain the direction h_F^k

Step 3 Compute the ratio ρ_k in (4)

Step 4 if $\rho_k \geqslant \eta_1$ and $\|g(x_F^k)\|^2 \geqslant \frac{\eta_2}{\mu^k}$, set $x_F^{k+1} \leftarrow x_F^k + h_F^k$ and $\mu^{k+1} = \max(\mu^k/\gamma, \mu_{\min})$; otherwise set $x_F^{k+1} = x_F^k$ and $\mu^{k+1} = \gamma\mu^k$. Then $k = k + 1$, go to step 1.

2 Convergence analysis

Assumption 1. *Suppose τ^k is the solution of* $\operatorname{argmin}_h m_k(x_F^k+h)$*, then the following condition hold:*

$$m_k(x_F^k) - m_k(x_F^k+\tau^k) \geqslant \frac{1}{4}\|g(x_F^k)\|^2 \min\left\{\frac{1}{\mu^k}, \frac{1}{\|H(x_F^k)\|}\right\} \tag{5}$$

and

$$\|\tau^k\| \leqslant \frac{2\|g(x_F^k)\|}{\mu^k} \tag{6}$$

Assumption 2. *Suppose $r_i(x)$ are continuously differentiable and $\nabla r_i(x)$ are Lipschitz continuous. $f(x)$ is bounded. $\|H(x)\| \leqslant c$ for a constant $c > 0$*

Moreover, Under Assumption 2., there exist a constant Lipschitz coffecient $L > 0$ and constrain $x_F, y_F \in F$ .Then the descent lemma tells

$$|f(y_F) - f(x_F) - \nabla f(x)^T (y-x)| \leqslant \frac{L}{2}\|y-x\|^2 \tag{7}$$

Lemma 3. *If Assumption 1. 2. holds. Then almost surely $\mu^k > \kappa$ for any $\kappa > 0$.*

Proof. For a constant $\kappa > 0$. Prove by contradiction. Assume the set $\{k \mid \mu^k < \kappa\}$ is infinite, also can conclude $P(\{k \mid \mu^k < \kappa\} = \infty) = \alpha > 0$. According to the algorithm, there has probability a that μ^k decrease infinite times. When the iteration is successful, $\rho_k \geqslant \eta_1$ and $\|g(x_F^k)\|^2 \geqslant \frac{\eta_2}{\mu^k}$ holds. Consider the set $S = \{k \mid \text{the } k\text{th iteration is successful}\}$

$$\sum_{k \in S} (f(x_F^k) - f(x_F^k + h_F^k)) \geqslant \sum_{k \in S} \frac{\eta_1}{4}\|g(x_F^k)\|^2 \min\left\{\frac{1}{\mu^k}, \frac{1}{\|H(x_F^k)\|}\right\} \geqslant \sum_{k \in S} \frac{\eta_1\eta_2}{4\kappa} \min\left\{\frac{1}{\kappa}, \frac{1}{c}\right\} \tag{8}$$

note that $|S| = \infty$ happens with positve probability α. So $E[\sum_{k \in S}(f(x_F^k) - f(x_F^k + h_F^k))] = \infty$. However, accoding to the assumption 2. $f(x)$ is bounded, which implies

$E[\sum_{k \in S}(f(x_F^k) - f(x_F^k + h_F^k))] \leqslant E[2f(x_F)] < \infty$. We obtain the contradiction with $P(\{k \mid \mu^k < \kappa\} = \infty) = \alpha > 0$. the proof is completed. □

Lemma 4. *If Assumption 1. 2. holds. When $\mu^k \geqslant \max\left\{c, \frac{8(L+c)}{1-\eta_1}\right\}$, then $\rho_k > \eta_1$.*

Proof. According to the Assumption 1. 2. and the condition $\mu^k \geqslant \max\left\{c, \frac{8(L+c)}{1-\eta_1}\right\}$, we derive the inequality

$$m_k(x_F^k) - m_k(x_F^k+h_F^k) \geqslant \frac{1}{4}\|g(x_F^k)\|^2 \min\left\{\frac{1}{\mu^k}, \frac{1}{\|H(x_F^k)\|}\right\} \tag{9}$$

recall the model $m_k(x_F^k) = f(x_F^k)$. Rewrite the the descent lemma

$$f(x_F^k + h_F^k) \leqslant m(x_F^k) + g(x_F^k)^T h_F^k + \frac{L}{2}\|h_F^k\|^2$$

thus,

$$f(x_F^k+h_F^k)-m_k(x_F^k+h_F^k)\leqslant\frac{L}{2}\|h_F^k\|^2-\frac{1}{2}h_F^{k\mathrm{T}}H(x_F^k)h_F^k-\frac{1}{2}\mu^k\|h_F^k\|^2\leqslant\frac{L+c}{2}\|h_F^k\|^2\leqslant 2(L+c)\frac{\|g(x_F^k)\|}{\mu^{k2}}$$

combined with the definition of ρ_k, $1-\rho_k\leqslant\frac{8(L+c)}{\mu^k}\Rightarrow\rho_k\geqslant 1-\frac{8(L+c)}{\mu^k}\geqslant\eta_1$. the proof is completed

□

The following lemmas prove Algorithm 1.1 convergence with probability. Set $\beta := d\sqrt{1-\frac{1}{2\mu^{k2}\|g(x^k)\|^4}}$

Lemma 5. *if b in $\mathcal{F}_b$ satisfies $b\geqslant\beta$, then the event*

$$I_k := \left\{\left|\|g(x_F^k)\|^2-\frac{b}{d}\|g(x^k)\|^2\right|<\frac{1}{\mu^k}\right\}$$

has $P(I_k)>\frac{1}{2}$.

Proof. Recall the definition $\mathcal{F}_b := \{F_l \mid F_l \subseteq$ the power set of $\{1,\cdots,d\}$, and $|F_l|=b\}$, so $|\mathcal{F}_b|=\binom{b}{d}$, and F is uniformly random chosen from $\mathcal{F}_b$.

$$E[\|g(x_F^k)\|^2]=E\left[\sum_{i\in F}(g_i(x^k))^2\right]=\sum_l\sum_{i\in F_l}(g_i(x^k))^2\,p(F_l)=\binom{b}{d}\frac{b}{d}\sum_{i=0}^{d}(g_i(x^k))^2\frac{1}{\binom{b}{d}}=\frac{b}{d}\|g(x^k)\|^2$$

and the variance

$$\begin{aligned}
\mathrm{Var}[\|g(x_F^k)\|^2] &= E[\|g(x_F^k)\|^4]-E[\|g(x_F^k)\|^2]^2\\
&= \sum_l\left(\sum_{i\in F_l}(g_i(x^k))^2\right)^2p(F_l)-\left(\frac{b}{d}\|g(x^k)\|^2\right)^2\\
&\leqslant \sum_l\left(\sum_{i=0}^{d}(g_i(x^k))^2\right)^2p(F_l)-\frac{b^2}{d^2}\|g(x^k)\|^4\\
&= \|g(x^k)\|^4-\frac{b^2}{d^2}\|g(x^k)\|^4\\
&\leqslant \left(1-\frac{\beta^2}{d^2}\right)\|g(x^k)\|^4
\end{aligned}$$

By the Chebyshev's inequality, we can obtain

$$P\left\{\left|\|g(x_F^k)\|^2-\frac{b}{d}\|g(x^k)\|^2\right|<\frac{1}{\mu^k}\right\}>1-\mu^{k2}\mathrm{Var}[\|g(x_F^k)\|^2]>\frac{1}{2}$$

□

Theorem 6. *Let the Assumption 1. 2. hold and condition in lemma 5 hold. Then the sequence of the total parameter* $\{x^k\}$ *generated by Algorithm, almost surely satisfies*

$$\lim_{k\to\infty}\inf\|g(x^k)\|=0$$

Proof. Prove this theorem by contradiction. Assume there exists $\varepsilon>0$ such that $\|g(x^k)\|^2\geqslant\frac{d}{b}\varepsilon$ for all $k\geqslant k_0$. According to the lemma 3., there exists $k>k_1$ such that

$$\mu^k>\chi:=\max\left\{\frac{2}{\varepsilon},\frac{2\eta_2}{\varepsilon},c,\frac{8(L+c)}{1-\eta_1},\gamma\mu_{\min}\right\} \tag{10}$$

Define $R_k=\log_\gamma\left(\frac{\mu^k}{\chi}\right)$, by the assumption, $R_k\leqslant 0$ for all $k>\max(k_0,k_1)$.

Since $\mu^k>\max\left\{c,\frac{8(L+c)}{1-\eta_1}\right\}$, then $\rho_k\geqslant\eta_1$. So the iteration success just depends on $\|g(x_F^k)\|^2$. In lemma 5., we have

$\left|\|g(x_F^k)\|^2-\frac{b}{d}\|g(x^k)\|^2\right|<\frac{1}{\mu^k}$ with probability $v>\frac{1}{2}$. $\left|\|g(x_F^k)\|^2-\frac{b}{d}\|g(x^k)\|^2\right|<\frac{1}{\mu^k}<\frac{\varepsilon}{2}$ then $\|g(x_F^k)\|^2\geqslant\frac{\varepsilon}{2}$. From (10), we can further obtain $\|g(x_F^k)\|^2>\frac{\eta_2}{\mu^k}$ which implies a successful iteration.

$$E[R_{k+1}]=v\left(\log_\gamma\left(\frac{\mu^k}{\chi\gamma}\right)\right)+(1-v)\log_\gamma\left(\frac{\mu^k\gamma}{\chi}\right)=v\left(\log_\gamma\left(\frac{\mu^k}{\chi}\right)-1\right)+(1-v)\left(\log_\gamma\left(\frac{\mu^k}{\chi}\right)+1\right)\geqslant R_k$$

Since $|R_{k+1}-R_k|\geqslant 1$, we can conclude $P[\lim_{k\to\infty}\sup R_k>0]=1$ which leads to a contradiction to our assumption: $R_k\leqslant 0$ for all $k>\max(k_0,k_1)$. So $\lim_{k\to\infty}\inf\|g(x^k)\|=0$ holds almost surely.

□