

Stochastic Optimization

Consider $\min_{x \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^N f_i(x)$

Example: Linear Regression for given data $(x_i, y_i)_{i=1}^N$
 $\phi_j(x)$ $j=1, 2, \dots, n$ are some model basis
Want to solve the eqn in least square sense.

$$c_1 \phi_1(x_i) + c_2 \phi_2(x_i) + \dots + c_n \phi_n(x_i) = y_i, \quad i=1, \dots, N$$

$g(x_i, c)$

$$\min_{c \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^N \|g(x_i, c) - y_i\|^2$$

(Full batch) Gradient Descent:

$$x_{k+1} = x_k - \eta \left[\frac{1}{N} \sum_{i=1}^N \nabla f_i(x) \right]$$

Stochastic Gradient Descent:

$$x_{k+1} = x_k - \eta_k \nabla f_{i(k)}(x_k)$$

$i(k) \in \{1, \dots, N\}$ are i.i.d. random variables
with uniform distribution

Now consider $\begin{cases} x_{k+1} = x_k - \eta_k A(x_k) \\ x_{k+1} = x_k - \eta_k A_{i(k)}(x_k) \end{cases}$

Example: $A_i = \nabla f_i$ and $A = \frac{1}{N} \sum_{i=1}^N \nabla f_i$
SGD GD

Def Monotone Operator

$$\langle Tx - Ty, x - y \rangle \geq 0$$

Example: $f(x)$ is convex, ∂f is monotone

$$\begin{cases} f(y) \geq f(x) + \langle \partial f(x), y - x \rangle \\ f(x) \geq f(y) + \langle \partial f(y), x - y \rangle \end{cases}$$

$$\Rightarrow \langle \partial f(x) - \partial f(y), x - y \rangle \geq 0$$

Def Graph of T is $\{(x, Tx), \forall x\}$

Def T is maximal monotone if its graph is not a strict subset of graph of a monotone operator S .

If $f(x)$ is well-defined & continuous, ∂f is maximal monotone

Assumptions:

- ① $A_i(x)$ are maximal monotone
- ② There are constants C_1, C_2 s.t.

$$\frac{1}{N} \sum_{i=1}^N \|A_i(x)\|^2 \leq \frac{C_1}{2} \|x\|^2 + C_2$$

Assume ∇f_i are L -cont.

$$\|\nabla f_i(x) - \nabla f_i(0)\| \leq L \|x - 0\|$$

$$\Rightarrow \|\nabla f_i(x)\| \leq L \|x\| + \|\nabla f_i(0)\|$$

- ③ $i(0), i(1), \dots, i(k), \dots$ are i.i.d. R.V.

with uniform distribution in $\{1, \dots, N\}$

$$\textcircled{3} \quad \sum_{k=0}^{\infty} \eta_k = \infty, \quad \sum_{k=0}^{\infty} \eta_k^2 < \infty$$

$$\eta_k = \frac{c}{(k+1)^p}, \quad k=0, 1, 2, \dots$$

$$\frac{1}{2} < p \leq 1$$

④ Assume A is demipositive
For any x_* satisfying $0 \in A(x_*)$

Example: $\langle Ax, x - x_* \rangle > 0, \quad \forall x$ and Ax does NOT include 0

2.2.3 Convergence for convex functions

Theorem 2.8. Assume $\nabla f(x)$ is Lipschitz-continuous with Lipschitz constant L and $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex. Then for any x, y :

1. $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2$
2. $\|\nabla f(x) - \nabla f(y)\|^2 \leq L \langle \nabla f(x) - \nabla f(y), x - y \rangle$.

1) Assume $f = \frac{1}{N} \sum_{i=1}^N f_i(x)$ is convex and $A = \nabla f$ is L -cont.

$$\langle \nabla f(x), x - x_* \rangle \geq f(x) - f(x_*) + \frac{1}{2L} \|\nabla f(x)\|^2$$

$$\geq \frac{1}{2L} \|\nabla f(x)\|^2 > 0$$

2) Assume $f(x)$ is convex, then $A = \partial f$ is demipositive

$$f(y) \geq f(x) + \langle \partial f(x), y - x \rangle$$

$$\Rightarrow f(x_*) \geq f(x) + \langle \partial f(x), x_* - x \rangle$$

$$\Rightarrow \langle \partial f(x), x - x_* \rangle \geq f(x) - f(x_*) > 0$$

Theorem 1

Under assumptions ①, ①, ②, ③ and demipositivity

for $x_{k+1} = x_k - \eta_k A_{i(k)}(x_k)$,
and any x_0 ,

$$x_k \rightarrow x_* \text{ with prob. 1}$$

Polyak-Ruppert Averaging:

For $x_{k+1} = x_k - \eta_k A_{i(k)}(x_k)$

define averaged iterate

$$\bar{x}_k = \frac{\sum_{j=0}^k \eta_j x_k}{\sum_{j=0}^k \eta_j}$$

Theorem 2 (Convergence with averaging)

Under assumptions ①, ①, ②, ③, without demipositivity

for $x_{k+1} = x_k - \eta_k A_{i(k)}(x_k)$ and any x_0 ,

$$\bar{x}_k \rightarrow x_* \text{ with prob. 1}$$

where $0 \in A(x_*)$

Proof of Theorem 1:

We make this general by also including $A = \partial f$ which is a set

$$x_{k+1} = x_k - \eta_k A_{i(k)}(x_k)$$

$$\begin{aligned} \Rightarrow \|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 + \eta_k^2 \|A_{i(k)}(x_k)\|^2 \\ &\quad - 2\eta_k \langle A_{i(k)}(x_k), x_k - x_* \rangle \end{aligned}$$

Assumption $\frac{1}{N} \sum_{i=1}^N \|A_i(x)\|^2 \leq \frac{C_1}{2} \|x\|^2 + C_2$

$$\begin{aligned} \Rightarrow \|A_{i(k)}(x_k)\|^2 &= \|A_{i(k)}(x_k) - A(x_*)\|^2 \\ &\leq 2\|A_{i(k)}(x_k)\|^2 + 2\|A(x_*)\|^2 \\ &\leq 2(\|x_k\|^2 + \|x_*\|^2) + 4C_2 \end{aligned}$$

$$\Rightarrow \|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 + \eta_k^2 \left[(\|x_k\|^2 + \|x_*\|^2) + 4C_2 \right]$$

$$- 2\eta_k \langle A_{i(k)}(x_k), x_k - x_* \rangle$$

Take conditional expectation E_k w.r.t. $i(k)$

for both sides,

$$\Rightarrow E_k \|x_k - x_*\|^2 \leq \|x_k - x_*\|^2$$

$$+ \eta_k^2 \left[(\|x_k\|^2 + \|x_*\|^2) + 4C_2 \right]$$

$$+ E_k \left[- 2\eta_k \langle A_{i(k)}(x_k), x_k - x_* \rangle \right]$$

||

$$- 2\eta_k \langle E_k(A_{i(k)}(x_k)), x_k - x_* \rangle$$

||

$$- 2\eta_k \langle \frac{1}{N} \sum_{i=1}^k A_i(x_k), x_k - x_* \rangle$$

||

$$- 2\eta_k \langle A(x_k), x_k - x_* \rangle$$

$$\Rightarrow E_k \|x_k - x_*\|^2 \leq \|x_k - x_*\|^2$$

$$+ \eta_k^2 \left[(\|x_k\|^2 + \|x_*\|^2) + 4C_2 \right]$$

$$- 2\eta_k \langle A(x_k), x_k - x_* \rangle$$

$$\|X_k\|^2 + \|X_*\|^2 \leq 2\|X_k - X_*\|^2 + 3\|X_*\|^2$$

$$E_k \|X_k - X_*\|^2 \leq (1 + 2\eta_k^2) \|X_k - X_*\|^2 + 3\eta_k^2 \left[\|X_*\|^2 + 4C_2 \right] \\ - 2\eta_k \langle A(X_k), X_k - X_* \rangle \\ \geq 0 \text{ due to demipositivity}$$

Theorem [Robbins & Siegmund Quasi-Martingale Convergence Theorem]

If

① $V_k \geq 0$, $S_k \geq 0$, $U_k \geq 0$ are R.V.

② $\sum_{k=0}^{\infty} \beta_k < \infty$, $\beta_k \geq 0$ are numbers

③ $E_k [V_{k+1}] \leq (1 + \beta_k) V_k - S_k + U_k$

④ $\sum_{i=1}^{\infty} U_i < +\infty$

Then

1) $V_k \rightarrow V_{\infty}$

2) $\sum_{k=0}^{\infty} S_k < +\infty$

with prob. 1

To apply this Quasi-Martingale Thm,

we need $\sum_{k=0}^{\infty} \eta_k^2 < +\infty$

everything is under prob. 1

$$\Rightarrow \textcircled{1} \sum_{k=0}^{\infty} \eta_k \langle A(X_k), X_k - X_* \rangle < \infty$$

$$\sum_{k=0}^{\infty} \eta_k = +\infty \quad \& \langle A(X_k), X_k - X_* \rangle \geq 0$$

$$\text{implies } \liminf_k \langle A(X_k), X_k - X_* \rangle = 0$$

$$\text{If } \liminf_k \langle A(X_k), X_k - X_* \rangle = c > 0,$$

$$\text{then } \sum \eta_k \langle A(X_k), X_k - X_* \rangle$$

$$\text{is essentially at least } \sum_{k=M}^{\infty} \eta_k c = +\infty$$

$$\textcircled{2} \|X_k - X_*\| \text{ has a limit}$$

$$\Rightarrow \{X_k\} \text{ is bounded (thus } A(X_k) \text{ bounded)}$$

$$\Rightarrow \text{There is a subsequence } X_{k_j} \text{ s.t.}$$

$$X_{k_j} \rightarrow X_{\infty}$$

$$A(X_{k_j}) \rightarrow A_{\infty}$$

$$\langle A(X_{k_j}), X_{k_j} - X_* \rangle \rightarrow 0$$

$$1) \liminf_k \langle A(X_k), X_k - X_* \rangle = 0$$

$$\Rightarrow \text{There is a subsequence } X_{k_n} \text{ s.t.}$$

$$\langle A(x_{k_n}), x_{k_n} - x^* \rangle \rightarrow 0$$

2) Pick another subsequence $x_{k_{n_i}}$ from x_{k_n}

$$\text{st. } A(x_{k_{n_i}}) \rightarrow a_\infty$$

3) Pick a subsequence $x_{k_{n_{i_j}}}$ from $x_{k_{n_i}}$

$$\text{st. } x_{k_{n_{i_j}}} \rightarrow x_\infty$$

So A is maximal monotone $\Rightarrow a_\infty \in A(x_\infty)$

$$\langle a_\infty, x_\infty - x^* \rangle = 0$$

$$\Rightarrow 0 \in \langle A(x_\infty), x_\infty - x^* \rangle$$

demi-positivity $\Rightarrow x_\infty$ is a zero point of A .