Stochastic Optimization

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

Example: Linear Regression for given data $(x_i, y_i)_{i=1}^{N}$

$\phi_j(x)$ $j=1,2,\cdots,n$ are some model basis

Want to solve the eqn in least square sense

$$\underbrace{c_1\phi_1(x_i)+c_2\phi_2(x_i)+\cdots+c_n\phi_n(x_i)}_{g(x_i,c)} = y_i \quad , \quad i=1,\cdots,N$$

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

(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:

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

$i(k)\in\{1,\cdots,N\}$ 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: $\underbrace{A_i = \nabla f_i}_{\text{SGD}}$ and $\underbrace{A = \frac{1}{N}\sum_{i=1}^{N}\nabla f_i}_{\text{GD}}$

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 $\{(x, T(x)), \forall x\}$

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

If $f(x)$ is a proper closed convex func, ∂f is maximal monotone

Assumptions:

⓪ $A_i(x)$ are maximal monotone

① 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.

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

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

② $i(0), i(1), \cdots, i(k), \cdots$ are i.i.d. R.V. with uniform distribution in $\{1, \cdots, N\}$

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

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

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

$\eta_k = \frac{1}{(k+1)^{\frac{1}{2}+\varepsilon}}$

④ Assume A is demipositive

For any x_* satisfying $0 \in A(x_*)$

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

2.2.3 Convergence for convex functions

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

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

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

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 $\left\{\begin{array}{l}\text{closed} \\ \text{convex} \\ \text{proper}\end{array}\right.$, 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 0, 1, 2, 3 and demipositivity

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

and any x_0,

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

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 0, 1, 2, 3, without demipositivity

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

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

where $0 \in A(x_*)$

Proof of Theorem 1:

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

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

$$\Rightarrow \|x_{k+1} - x_*\|^2 = \|x_k - x_*\|^2 + \eta_k^2 \|A_{i(k)}(x_k)\|^2$$

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

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

$$\Rightarrow \|A_{i(k)}(x_k)\|^2 = \|A_{i(k)}(x_k) - A(x_*)\|^2$$

$$\leq 2\|A_{i(k)}(x_k)\|^2 + 2\|A(x_*)\|^2$$

$$\leq 2\left(\|x_k\|^2 + \|x_*\|^2\right) + 4C_2$$

$$\Rightarrow \|x_{k+1} - x_*\|^2 \le \|x_k - x_*\|^2 + \eta_k^2\left[2\left(\|x_k\|^2 + \|x_*\|^2\right) + 4C_2\right] - 2\eta_k \langle A_{i(k)}(x_k), x_k - x_* \rangle$$

Take conditional expection E_k w.r.t. $i(k)$ for both sides,

$$\Rightarrow E_k \|x_k - x_*\|^2 \le \|x_k - x_*\|^2 + \eta_k^2\left[2\left(\|x_k\|^2 + \|x_*\|^2\right) + 4C_2\right] + \underbrace{E_k\left[-2\eta_k \langle A_{i(k)}(x_k), x_k - x_* \rangle\right]}$$

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

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

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

$$\Rightarrow E_k \|x_k - x_*\|^2 \le \|x_k - x_*\|^2 + \eta_k^2\left[2\left(\|x_k\|^2 + \|x_*\|^2\right) + 4C_2\right] - 2\eta_k \langle A(x_k), x_k - x_* \rangle$$

$$\|x_k\|^2 + \|x_*\|^2 \leq 2\|x_k - x_*\|^2 + 3\|x_*\|^2$$

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

<u>Theorem</u> [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) $V_k \to V_\infty$

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

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

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

implies $\liminf_k \langle A(x_k), x_k - x_* \rangle = 0$

If $\liminf_k \langle A(x_k), x_k - x_* \rangle = C > 0$,

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

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

② $\| x_k - x_* \|$ has a limit

$\| \nabla f(x_k) - \nabla f(x_*) \| \leq L \| x_k - x_* \|$

$\Rightarrow$ $\{x_k\}$ is bounded (thus $A(x_k)$ bounded)

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

$x_{k_j} \to x_\infty$

$A(x_{k_j}) \to a_\infty$

$\langle A(x_{k_j}), x_{k_j} - x_* \rangle \to 0$

1) $\liminf_k \langle A(x_k), x_k - x_* \rangle = 0$

$\Rightarrow$ There is a subsequene x_{k_n} s.t.

$$\langle A(x_{k_n}), x_{k_n} - x_* \rangle \to 0$$

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

s.t. $A(x_{k_{n_i}}) \to a_\infty$

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

s.t. $x_{k_{n_{i_j}}} \to x_\infty$

So $\left\{ \begin{array}{l} A \text{ is maximal monotone} \Rightarrow a_\infty \in A(x_\infty) \\ \langle a_\infty, x_\infty - x_* \rangle = 0 \end{array} \right.$

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

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

Example:

① $\lim_j \nabla f(x_{k_j}) = \nabla f(\lim_j x_{k_j})$

② $\langle \lim_j \nabla f(x_{k_j}), \lim_j x_{k_j} - x_* \rangle = 0$

③ $\langle \nabla f(x_\infty), x_\infty - x_* \rangle = 0$

Assume $\begin{cases} \text{convexity of } f(x) \\ L\text{-cont. of } \nabla f \end{cases}$

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

So $\langle \nabla f(x_\infty), x_\infty - x_* \rangle = 0$

$\Rightarrow \|\nabla f(x_\infty)\| = 0 \Rightarrow x_\infty$ is a minimizer