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 $(x_i, y_i)_{i=1}^{N}$

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

Want to solve the eqn in least square sense.

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

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

(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

---

I. Convergence of Gradient Flow

$$\begin{cases} x'(t) = -\nabla f[x(t)] \\ x(0) = x_0 \end{cases}$$

Three steps:

① Pick / Construct a Lyapunov function $\mathcal{L}$

② Show $\mathcal{L}$ converges for $t \to \infty$

③ Show the limit gives what we want

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

$y = x_* \Rightarrow \quad \langle \nabla f(x), x - x_* \rangle \geq \frac{1}{L} \|\nabla f(x)\|^2 \geq 0$

Proof for gradient flow :

① $\mathcal{L}(t) = \|x(t) - x_*\|^2 \geq 0$

② $\mathcal{L}'(t) = 2 \langle x(t) - x_*, x'(t) \rangle$

$= 2 \langle x(t) - x_*, -\nabla f[x(t)] \rangle \leq -\frac{2}{L} \|\nabla f(x)\|^2$

$\Rightarrow \int_0^\infty \mathcal{L}'(t)\, dt \leq -\frac{2}{L} \int_0^\infty \|\nabla f[x(t)]\|^2 \, dt$

[$\mathcal{L}'(t) < 0 \Rightarrow \mathcal{L}(t) \searrow \Rightarrow \mathcal{L}_\infty = \lim_{t \to \infty} \mathcal{L}(t)$ exists]

$\Rightarrow \mathcal{L}_\infty - \mathcal{L}(0) \leq -\frac{2}{L} \int_0^\infty \|\nabla f[x(t)]\|^2 \, dt$

$\Rightarrow \int_0^\infty \|\nabla f(x)\|^2 \, dt \leq \frac{L}{2} \mathcal{L}(0) < \infty$

$\Rightarrow$ for any n
there is $t_n \geq n$ s.t. $\|\nabla f[x(t_n)]\|^2 < \frac{1}{n^2}$

because otherwise $\|\nabla f[x(t)]\|^2 \geq \frac{1}{n^2}$ for any $t \geq n$

$\Rightarrow \int_0^\infty \|\nabla f[x(t)]\|^2 dt \geq \int_n^\infty \|\nabla f[x(t)]\|^2 dt \geq \infty$

$\Rightarrow \nabla f[x(t_n)] \to 0$, $n \to \infty$

③ $x(t_n)$ is bounded because $\lim_{t\to\infty} \|x(t) - x_*\|^2 < \infty$

$\Rightarrow$ convergent subsequence $x(t_{n_j}) \to \bar{x}$

Continuity of $\nabla f \Rightarrow \nabla f[x(t_{n_j})] \to \nabla f(\bar{x})$

$\Rightarrow \nabla f(\bar{x}) = 0$

$\Rightarrow \bar{x}$ is critical point of f

$\Rightarrow \bar{x} \in \operatorname{argmin}(f)$

Use $x_* = \bar{x}$ in ①

$\Rightarrow \mathcal{L}(t) = \|x(t) - \bar{x}\|^2 \searrow$

$\forall t > t_{n_j}$, $\|x(t) - \bar{x}\| \leq \|x(t_{n_j}) - \bar{x}\|$

$\Rightarrow x(t) \to \bar{x}$.

II. Convergence of Gradient Descent

Assume $\begin{cases} f(x) \text{ has at least one minimizer} \\ f(x) \text{ is convex \& } \nabla f \text{ is } L\text{-cont.} \end{cases}$

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

$y = x_* \Rightarrow \quad \langle \nabla f(x), x - x_* \rangle \geq \frac{1}{L}\|\nabla f(x)\|^2 \geq 0$

Proof

Step I For a fixed x_*, define $\mathcal{L}_k = \|x_k - x_*\|^2$

Step II

$$\|x_{k+1} - x_*\|^2 = \|x_k - \eta \nabla f(x_k) - x_*\|^2$$

$$= \|x_k - x_*\|^2 - 2\eta \langle \nabla f(x_k), x_k - x_* \rangle + \eta^2 \|\nabla f(x_k)\|^2$$

$$\leq \|x_k - x_*\|^2 - \eta\left(\frac{2}{L} - \eta\right)\|\nabla f(x_k)\|^2$$

$$\mathcal{L}_{k+1} \leq \mathcal{L}_k - \eta\left(\frac{2}{L} - \eta\right)\|\nabla f(x_k)\|^2$$

$0 < \eta < \frac{2}{L} \Rightarrow \mathcal{L}_k \searrow$

$$\Rightarrow \sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2 \leq \frac{1}{\eta\left(\frac{2}{L} - \eta\right)} \mathcal{L}_0 < \infty$$

$$\Rightarrow \quad \nabla f(x_k) \to 0 \qquad (*)$$

Step III $\mathcal{L}_k = \|x_k - x_*\|^2 \searrow$

$\Rightarrow \{x_k\}$ is bounded

$\Rightarrow$ convergent subseq. $x_{k_j} \to \bar{x}$

$\nabla f(x)$ is cont.

$\Rightarrow \nabla f(x_{k_j}) \to \nabla f(\bar{x})$

$(*) \Rightarrow \nabla f(\bar{x}) = 0 \Rightarrow \bar{x} \in \operatorname{argmin}(f)$

Use $x_* = \bar{x}$ in Step I & II $\Rightarrow$

$\forall\, k > k_j$, $\|x_k - \bar{x}\| \leq \|x_{k_j} - \bar{x}\|$

$\Rightarrow x_k \to \bar{x}$

III. Convergence of Stochastic Gradient Descent

(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

---

Take conditional expection E_k w.r.t. $i(k)$ conditioned on $i(k-1), \cdots, i(1), i(0)$

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x) \Rightarrow \nabla f = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x)$$

$$E_k\left[\nabla f_{i(k)}(x_k)\right] = \sum_{i=1}^{N} \frac{1}{N} \nabla f_i(x_k)$$

$$E_k \|\nabla f_{i(k)}(x_k)\|^2 = \sum_{i=1}^{N} \frac{1}{N} \|\nabla f_i(x_k)\|^2$$

Assume { For simplicity, let's assume X_* is unique
$f(x)$ is convex & ∇f is L-cont.
each $f_i(x)$ is L-Cont.

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

$y = x_* \Rightarrow \quad \langle \nabla f(x), x - x_* \rangle \geq \frac{1}{L}\|\nabla f(x)\|^2 \geq 0$

Proof

Step I For a fixed x_*, define $\mathcal{L}_k = \|x_k - x_*\|^2$

Step II

$$\|x_{k+1} - x_*\|^2 = \|x_k - \eta \nabla f_{i(k)}(x_k) - x_*\|^2$$

$$= \|x_k - x_*\|^2 - 2\eta \langle \nabla f_{i(k)}(x_k), x_k - x_* \rangle + \eta^2 \|\nabla f_{i(k)}(x_k)\|^2$$

Take E_k for both sides

$$E_k\|x_{k+1} - x_*\|^2 = E_k\|x_k - x_*\|^2 - E_k 2\eta \langle \nabla f_{i(k)}(x_k), x_k - x_* \rangle + \eta^2 E_k \|\nabla f_{i(k)}(x_k)\|^2$$

$$= \|x_k - x_*\|^2 - 2\eta \langle E_k \nabla f_{i(k)}(x_k), x_k - x_* \rangle + \eta^2 E_k \|\nabla f_{i(k)}(x_k)\|^2$$

$$= \|x_k - x_*\|^2 - 2\eta \langle \nabla f(x_k), x_k - x_* \rangle + \eta^2 \sum_{i=1}^{N} \frac{1}{N} \|\nabla f_i(x_k)\|^2$$

$$\left(\langle \nabla f(x), x - x_* \rangle \geq \frac{1}{L}\|\nabla f(x)\|^2 \geq 0 \right)$$

$$\leq \|x_k - x_*\|^2 - 2\eta \frac{1}{L} \|\nabla f(x_k)\|^2 + \eta^2 \sum_{i=1}^{N} \frac{1}{N} \|\nabla f_i(x_k)\|^2$$

$$\leq \|x_k - x_*\|^2 - 2\eta \frac{1}{L} \|\nabla f(x_k)\|^2 + 2\eta^2 L^2 \|x_k - x_*\|^2 + A\eta^2$$

Assume $f_i(x)$ is L-cont. $\Rightarrow$

$$A = \frac{2}{N}\sum_{i=1}^{N} \|\nabla f_i(x_*)\|^2$$

$$\begin{aligned}\|\nabla f_i(x)\|^2 &= \|\nabla f_i(x) - \nabla f_i(x_*) + \nabla f_i(x_*)\|^2 \\ &\le 2\|\nabla f_i(x) - \nabla f_i(x_*)\|^2 + 2\|\nabla f_i(x_*)\|^2 \\ &\le 2L^2\|x - x_*\|^2 + 2\|\nabla f_i(x_*)\|^2\end{aligned}$$

$$\mathcal{L}_k = \|x_k - x_*\|^2$$

$$E_k\,\mathcal{L}_{k+1} \le (1 + 2L^2\eta_k^2)\,\mathcal{L}_k - \frac{2\eta_k}{L}\|\nabla f(x_k)\|^2 + A\eta_k^2$$

Theorem [Robbins & Siegmund Quasi-Martingale Convergence Theorem]

If

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

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

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

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

$$E_k\,\mathcal{L}_{k+1} \le (1 + 2L^2\eta_k^2)\,\mathcal{L}_k - \frac{2\eta_k}{L}\|\nabla f(x_k)\|^2 + A\eta_k^2$$

$$\beta_k = 2L^2\eta_k^2 \qquad S_k = \frac{2}{L}\eta_k\|\nabla f(x_k)\|^2 \qquad U_k = A\eta_k^2$$

$\sum_{i=1}^{\infty} \eta_i^2 < \infty \Rightarrow$ ② & ④ are OK

Quasi-Martingale Thm $\Rightarrow$ (with prob. 1)

(In other words, there is probability event Ω with $P(\Omega)=1$, (*) holds on Ω)

(*) $\begin{cases} \mathcal{L}_k \to \mathcal{L}_\infty \\ \sum_{k=0}^{\infty} \eta_k \|\nabla f(x_k)\|^2 < \infty \end{cases}$

Step III $P(\Omega)=1$ and for any $\omega \in \Omega$

(*) holds ω is one particular choice of $i(0), i(1), \cdots, i(k), \cdots$

For any fixed $\omega \in \Omega$,

assume $\sum_{k=1}^{\infty} \eta_k = \infty$,

$\sum_{k=1}^{\infty} \eta_k \|\nabla f(x_k)\|^2 < \infty \Rightarrow \liminf_{k\to\infty} \|\nabla f(x_k)\|^2 = 0$

$\Rightarrow \nabla f(x_{k_j}) \to 0$

x_{k_j} is bounded $\Rightarrow$ a subsequence of it converges

still use x_{k_j} for it

$\Rightarrow \quad x_{k_j} \to \bar{x} \Rightarrow \nabla f(x_{k_j}) \to \nabla f(\bar{x})$

$\Rightarrow \bar{x} = x_*$

$\Rightarrow L_\infty = 0$

$L_k \to L_\infty \Rightarrow L_k \to 0$

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

$i(k)$ is i.i.d. uniform

Assume
- $f(x)$ has minimizers
- $f(x)$ is convex & ∇f is L-cont.
- each $f_i(x)$ is L-cont.
- $\sum_{k=1}^{\infty} \eta_k = \infty$, $\sum_{k=1}^{\infty} \eta_k^2 < \infty$

then with prob. 1, $x_k \to x_*$ (one of minimizers)