Notes for Optimization

Jiyao Liu
CIS Department (of Temple University)
Philadelphia, USA
jiyao.liu@temple.edu

***Abstract*—This note discusses optimization in federated learning, including properties of loss functions and some basic equations and inequalities.**

I. Assumptions

A. Smoothness

If function f is L-smooth, then, for $\forall x, \forall y \in \mathbb{R}^d, L > 0$,

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

$$\langle \nabla f(x) - \nabla f(y), x-y \rangle \le L\|x-y\|^2 \tag{2}$$

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\| \tag{3}$$

B. Convexity

If function f is μ-convex, then, for $\forall x, \forall y \in \mathbb{R}^d, \mu > 0$,

$$f(y) \ge f(x) + \langle \nabla f(x), (y-x) \rangle + \frac{\mu}{2}\|x-y\|^2 \tag{4}$$

$$\langle \nabla f(x) - \nabla f(y), x-y \rangle \ge \mu\|x-y\|^2 \tag{5}$$

$$\|\nabla f(x) - \nabla f(y)\| \ge \mu\|x-y\| \tag{6}$$

$$\alpha f(x) + \beta f(y) \le f(\alpha x + \beta y), \alpha + \beta = 1 \tag{7}$$

C. Bounded Gradient

If the datasets on all devices are IID, then we can assume, for any device i, loss function F_i, parameters $\mathbf{x}$, and dataset ξ, there exists $G > 0$,

$$\|\nabla F_i(\mathbf{x}; \xi)\|^2 \le G^2 \tag{8}$$

D. Bounded Variance

Bounded stochastic gradient variance

$$f_i(\mathbf{x}) \triangleq \mathbb{E}[F_i(\mathbf{x})]\|^2 \tag{9}$$

$$\mathbb{E}_{\xi \sim \mathcal{D}_i}[\|\nabla F_i(\mathbf{x}; \xi) - \nabla f_i(\mathbf{x})] \le \sigma^2 \tag{10}$$

E. Bounded Dissimilarity

If the datasets on all devices are non-IID, we can assume, for all N devices, loss function f_i of device i, and any parameters $\mathbf{x}$, there exists $\kappa > 0$,

$$\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le \zeta^2 \tag{11}$$

II. Equations

A. Expectation of Squared Norm

For $\forall \mathbf{v} \in \mathbb{R}^n$,

$$\mathbb{E}[\|\mathbf{v}\|^2] = \mathbb{E}[\|\mathbf{v} - \mathbb{E}[\mathbf{v}]\|^2] + \|\mathbb{E}[\mathbf{v}]\|^2$$

Proof. See section IV-A.

B. Parallelogram Law

$\forall \mathbf{u}, \forall \mathbf{v} \in \mathbb{R}^n$,

$$\|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2 \tag{12}$$

C. Theorem 1

For $\forall a, \forall b, \forall c \in \mathbb{R}^n$,

$$2\langle a-b, a-c \rangle = \|a-b\|^2 + \|a-c\|^2 - \|b-c\|^2$$

Proof.

$$\begin{aligned}
&2\langle a-b, a-c \rangle \\
&= \langle a-b, a-c \rangle + \langle a-b, a-c \rangle \\
&= \langle a-c-(b-c), a-c \rangle + \langle a-b, a-b+(b-c) \rangle \\
&= \langle a-c, a-c \rangle + \langle b-c, c-a \rangle \\
&\quad + \langle a-b, a-b \rangle + \langle a-b, b-c \rangle \\
&= \langle a-c, a-c \rangle - \langle b-c, b-c \rangle + \langle a-b, a-b \rangle \\
&= \|a-b\|^2 + \|a-c\|^2 - \|b-c\|^2
\end{aligned}$$

III. Inequalities

A. Sum in Norm Expansion

$$\|\sum_i v_i\|^2 \le \sum_i \|v_i\|^2$$

This is easy to prove by hand so we omit the proof here.

B. Cauchy-Schwarz Inequality

Cauchy-Schwarz inequality in $\mathbb{R}^n$ states that for $\forall \mathbf{u}, \forall \mathbf{v} \in \mathbb{R}^n$,

$$\|\langle \mathbf{u}, \mathbf{v} \rangle\|^2 \le \|\mathbf{u}\|^2\|\mathbf{v}\|^2$$

Proof. See section IV-C.

C. AM-GM Inequality

AM-GM inequality in $\mathbb{R}^n$ states that for $\forall \mathbf{u}, \forall \mathbf{v} \in \mathbb{R}^n$,

D. Young's Inequality

$\forall \mathbf{u}, \forall \mathbf{v} \in \mathbb{R}^n$, $p > 1$,

$$\mathbf{uv} \le \frac{p-1}{p}\mathbf{u}^{\frac{p}{p-1}} + \frac{1}{p}\mathbf{v}^p$$

When $p = 2$,

$$2\mathbf{uv} \le \mathbf{u}^2 + \mathbf{v}^2$$

Proof: link (only for real numbers currently).

References

[1] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

IV. Appendix

A. Proof of Expectation of Squared Norm

Here we prove

$$\mathbb{E}[\|\mathbf{v}\|^2] = \mathbb{E}[\|\mathbf{v} - \mathbb{E}[\mathbf{v}]\|^2] + \|\mathbb{E}[\mathbf{v}]\|^2$$

According to the definition of $\|\cdot\|$,

$$\mathbb{E}[\|\mathbf{v}\|^2] = \mathbb{E}[\sum_i v_i^2] \tag{13}$$

$$\|\mathbb{E}[\mathbf{v}]\|^2 = \sum_i (\mathbb{E}[v_i])^2 \tag{14}$$

Then,

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{v} - \mathbb{E}[\mathbf{v}]\|^2] \\
&= \mathbb{E}[\sum_i (v_i - \mathbb{E}[v_i])^2] \\
&= \mathbb{E}[\sum_i v_i^2 - 2\sum_i v_i \mathbb{E}[v_i] + \sum_i (\mathbb{E}[v_i])^2] \\
&= \mathbb{E}[\sum_i v_i^2] - 2\mathbb{E}[\sum_i v_i \mathbb{E}[v_i]] + \mathbb{E}[\sum_i (\mathbb{E}[v_i])^2] \\
&= \mathbb{E}[\sum_i v_i^2] - 2\sum_i (\mathbb{E}[v_i])^2 + \sum_i (\mathbb{E}[v_i])^2 \\
&= \mathbb{E}[\sum_i v_i^2] - \sum_i (\mathbb{E}[v_i])^2 \qquad (15)
\end{aligned}$$

Combine (13), (14), (15), we get the result.

B. Proof of Parallelogram Law

The equation is equivalent to

$$\sum_i (u_i + v_i)^2 + \sum_i (u_i - v_i)^2 = 2\sum_i u_i^2 + 2\sum_i v_i^2$$

$$\sum_i 2(u_i^2 + v_i^2) = 2\sum_i u_i^2 + 2\sum_i v_i^2$$

C. Proof of Cauchy-Schwarz Inequality

Proof. It is equivalent to prove

$$\left(\sum_{i=1}^n u_i v_i\right)^2 \leq \left(\sum_{i=1}^n u_i^2\right)\left(\sum_{i=1}^n v_i^2\right).$$

Consider

$$\sum_{i=1}^n (u_i x + v_i)^2 \geq 0$$

$$\left(\sum_{i=1}^n u_i^2\right) x^2 + 2\left(\sum_{i=1}^n u_i v_i\right) x + \sum_{i=1}^n v_i^2 \geq 0$$

This quadratic polynomial in x has at most 1 real root, thus, its discriminant $\Delta \leq 0$. That is

$$4\left(\sum_{i=1}^n u_i v_i\right) - 4\left(\sum_{i=1}^n u_i^2\right)\left(\sum_{i=1}^n v_i^2\right) \leq 0$$

$$\left(\sum_{i=1}^n u_i v_i\right)^2 \leq \left(\sum_{i=1}^n u_i^2\right)\left(\sum_{i=1}^n v_i^2\right).$$