

Machine Learning Theory: Introduction

Richard Xu

August 16, 2021

1 Something different: Gradient descend convergence for β -smooth function

You need three definitions and theorems for β -smooth function:

1. **definition** of convex function:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad (1)$$

2. **definition** of β -smooth convex function requires:

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|^2 \quad (2)$$

QUESTION: what kind of function is $\nabla f(\cdot)$?

3. **Theorem 1** if convex function f is β -smooth, then for gradient descend $x_{t+1} = x_t - \eta \nabla f(x_t)$ and when $\eta = \frac{1}{\beta}$, we have:

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|_2^2 \quad (3)$$

QUESTION: what does it tell you about β -smooth convex function guarantees?

Theorem 2 if convex function f is β -smooth, for gradient descend $x_{t+1} = x_t - \eta \nabla f(x_t)$, with learning rate $\eta = \frac{1}{\beta}$, then:

$$\begin{aligned} \implies \epsilon \equiv f(x_T) - f(x^*) &\leq \frac{1}{T} \sum_{t=0}^{T-1} f(x_{t+1}) - f(x^*) \\ &\leq \frac{1}{T} \frac{1}{2\eta} \left(\|x_0 - x^*\|_2^2 - \|x_T - x^*\|_2^2 \right) \end{aligned} \quad (4)$$

1. which means $\epsilon(t) = O(\frac{1}{t})$, in word, it says it takes $t \times$ “some constant” iterations to achieve error $\frac{1}{\epsilon}$
2. first inequality line is due to **Theorem 1**

The **proof** began by Eq.(3), conditioned on $\eta = \frac{1}{\beta}$:

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) - \frac{\eta}{2} \|\nabla f(x_t)\|_2^2 \\ &\leq \textcolor{red}{f(x^*)} + \langle \nabla f(x_t), x_t - x^* \rangle - \frac{\eta}{2} \|\nabla f(x_t)\|_2^2 \quad \text{we need to bring in } x^* \end{aligned} \quad (5)$$

QUESTION: how do you get the **red** part?

looking at Eq.(5), we need to expand $\langle \nabla f(x_t), x_t - x^* \rangle$. The **first attempt** is:

$$\begin{aligned} \|a - b\|_2^2 &= \|a\|_2^2 - 2\langle a, b \rangle + \|b\|_2^2 \\ \implies \langle a, b \rangle &= \frac{\|a\|_2^2 + \|b\|_2^2 - \|a - b\|_2^2}{2} \end{aligned} \quad (7)$$

so by letting $a = x_t - x^*$, and $b = \nabla f(x_t)$, we have:

the above still not very useful, so **second attempt** is to make a little modification, where $a \rightarrow a$, and $b \rightarrow \eta b$

$$\begin{aligned} \|a - \eta b\|_2^2 &= \|a\|_2^2 - 2\langle a, \eta b \rangle + \|\eta b\|_2^2 \\ \implies \eta \langle a, b \rangle &= \frac{\|a\|_2^2 + \|\eta b\|_2^2 - \|a - \eta b\|_2^2}{2} \\ \implies \langle a, b \rangle &= \frac{\|a\|_2^2 + \|\eta b\|_2^2 - \|a - \eta b\|_2^2}{2\eta} \end{aligned} \quad (9)$$

instead, i.e., $a = x_t - x^*$, and $\eta b = \eta \nabla f(x_t)$, brings nice cancellation:

finally:

$$\begin{aligned}
\epsilon \equiv f(x_T) - f(x^*) &\leq \frac{1}{T} \sum_{t=0}^{T-1} f(x_{t+1}) - f(x^*) = \frac{1}{T} \sum_{t=1}^{T-1} \frac{1}{2\eta} \left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \right) \\
&= \frac{1}{T} \frac{1}{2\eta} \left(\|x_0 - x^*\|_2^2 - \|x_T - x^*\|_2^2 \right)
\end{aligned} \tag{11}$$

QUESTION: why can we conclude $f(x_T) - f(x^*) \leq \frac{1}{T} \sum_{t=0}^{T-1} f(x_{t+1}) - f(x^*)$?

2 what concentration inequality is about?

imagine we have an “exact” (but un-achievable) bound, i.e., concentration equality

$$\begin{aligned} \Pr(X > \epsilon) &= \beta \quad \text{or} \\ \Pr(X < \epsilon) &= 1 - \beta \quad \text{or} \end{aligned} \tag{12}$$

then, let’s look at concentration inequality:

3 To bound a positive random variable: Markov Inequality

Theorem 3 if X has support in \mathcal{R}^+ :

$$\Pr(X \geq s) \leq \frac{\mathbb{E}(X)}{s} \quad (13)$$

proof for this can be understood by:

$$\begin{aligned} \mathbb{E}(X) &= \int_{-\infty}^{\infty} xp(x) \, dx = \int_0^{\infty} xp(x) \, dx \quad \text{since } x > 0 \\ &= \int_0^s xp(x) \, dx + \int_s^{\infty} xp(x) \, dx \quad \text{pick arbitrary bound } s \\ &\geq \int_s^{\infty} xp(x) \, dx \quad x > 0 \implies \int_0^s xp(x) \, dx > 0 \\ &\geq \int_s^{\infty} sp(x) \, dx \\ &= s \int_s^{\infty} p(x) \, dx \\ &= s\Pr(X \geq s) \\ \implies \Pr(X \geq s) &\leq \frac{\mathbb{E}(X)}{s} \end{aligned} \quad (14)$$

3.1 second proof

$$\begin{cases} \text{when } X < a : \mathbb{1}_{(X \geq a)} = 0 \implies a\mathbb{1}_{(X \geq a)} = \underbrace{0 \leq X}_{\text{due to support of } X \geq 0} \\ \text{when } X \geq a : \mathbb{1}_{(X \geq a)} = 1 \implies a\mathbb{1}_{(X \geq a)} = \underbrace{a \leq X}_{\text{due to condition } (X \geq a)} \end{cases} \quad (15)$$

in both cases, we have: $a\mathbb{1}_{(X \geq a)} \leq X$, then:

4 Chebyshev's inequality

Chebyshev's inequality is the absolute version of Tail bound, as oppose to Chernoff bound (without absolute value):

$$\begin{aligned}\Pr(|X - \mathbb{E}(X)| \geq \epsilon) &= \Pr((X - \mathbb{E}(X))^2 \geq \epsilon^2) \\ &\leq \frac{\mathbb{E}[(X - \mathbb{E}(X))^2]}{\epsilon^2} \\ &= \frac{\text{Var}(X)}{\epsilon^2}\end{aligned}\tag{17}$$

4.1 Useful Fact

1. $\Pr(|X - \mathbb{E}(X)| \geq \epsilon) = \Pr((X - \mathbb{E}(X))^2 \geq \epsilon^2)$, so you do not need to deal with $|\cdot|$. This fact can be used generically.
2. Although it's obvious, but only if you can prove symmetry, i.e.,:

$$\begin{aligned}\Pr(X - \mathbb{E}[X] \geq \epsilon) &\leq C \\ \Pr(\mathbb{E}[X] - X \geq -\epsilon) &\leq C\end{aligned}\tag{18}$$

then you can claim:

$$\Pr(|X - \mathbb{E}(X)| \geq \epsilon) \leq 2C\tag{19}$$

4.2 alternative expressions of Chebyshev's inequality

$$\begin{aligned}\Pr(|X - \mathbb{E}(X)| \geq \epsilon) &\leq \frac{\text{Var}(X)}{\epsilon^2} \\ \implies \Pr\left(\left|\frac{X - \mathbb{E}(X)}{\sigma(X)}\right| \geq \epsilon\right) &\leq \frac{1}{\epsilon^2} \quad \text{standardize R.V. s.t. its variance is 1} \\ \implies \Pr(|X - \mathbb{E}(X)| \geq \epsilon\sigma(X)) &\leq \frac{1}{\epsilon^2}\end{aligned}\tag{20}$$

4.3 application of Chebyshev's inequality

we can use it to derive weak law of large number:

$$\forall \epsilon > 0 \quad \lim_{n \rightarrow \infty} \Pr(|\bar{X}_n - \mu| \leq \epsilon) = 1\tag{21}$$

this means that $\bar{X}_n \xrightarrow{P} \mu$, as $n \rightarrow \infty$, i.e., \bar{X}_n converge in probability to μ as $n \rightarrow \infty$

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\ &= \left(\frac{1}{n}\right)^2 \text{var}(X_1 + X_2 + \dots + X_n) \\ &= \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \dots + \sigma^2) \quad (\text{since the } X_i\text{'s are independent}) \\ &= \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}\end{aligned}\tag{22}$$

using **Chebyshev's inequality**, let:

this means that $\bar{X}_n \xrightarrow{P} \mathbb{E}(\bar{X})$, as $n \rightarrow \infty$, i.e., \bar{X}_n converge in probability to $\mathbb{E}(\bar{X})$ as $n \rightarrow \infty$
note that the tail probability $\Pr(|X - \mathbb{E}(X)| \geq \epsilon)$ is decaying $O(\frac{1}{n})$ so it's actually quite slow.

5 Different types of Convergence

$$X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{d} X \quad (24)$$

5.1 Convergence in probability

these are equivalent:

$$\begin{aligned} & X_n \xrightarrow{P} X \\ & \lim_{n \rightarrow \infty} \Pr(|X_n - X| \leq \epsilon) = 1, \quad \forall \epsilon \\ & \forall \epsilon, \delta \quad \exists N_{\epsilon, \delta} \quad \text{s.t. } P(|X_n - X| \geq \epsilon) \leq \delta \quad \forall n > N_{\epsilon, \delta} \quad \text{note there is no limit} \\ & X_n = o_p(1) \end{aligned} \quad (25)$$

5.2 Example of $X_n \xrightarrow{d} X$: Central Limit Theorem

Theorem 4 if X_i (**any arbitrary distribution**) has finite non-zero variance σ^2 , for large n , \bar{X}_n approximately has a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

We can put **Theorem(4)** in equation:

$$\lim_{n \rightarrow \infty} \Pr\left(a \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq b\right) = \Phi(b) - \Phi(a) \quad (26)$$

QUESTION: what type of convergence is above?

5.3 Show CLT implies WLLN

Let $a = -\frac{c}{\sigma}$ and $b = \frac{c}{\sigma}$, and use Eq.(26)

$$\begin{aligned} & \lim_{n \rightarrow \infty} \Pr\left(-\frac{c}{\sigma} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq \frac{c}{\sigma}\right) = \Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right) \\ & \lim_{n \rightarrow \infty} \Pr(-c \leq \sqrt{n}(\bar{X}_n - \mu) \leq c) = \Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right) \\ & \lim_{n \rightarrow \infty} \Pr\left(-\frac{c}{\sqrt{n}} \leq \bar{X}_n - \mu \leq \frac{c}{\sqrt{n}}\right) = \Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right) \\ & \lim_{n \rightarrow \infty} \Pr\left(|\bar{X}_n - \mu| \leq \frac{c}{\sqrt{n}}\right) = \Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right) \end{aligned} \quad (27)$$

1. As n becomes larger, we have a threshold $N_{\epsilon, \delta}$ such that, $\forall n > N_{\epsilon, \delta}$:

$$\Pr\left(|\bar{X}_n - \mu| \leq \frac{c}{\sqrt{n}}\right) \geq \Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right) - \frac{\delta}{2} \quad (28)$$

2. $\forall n \geq N_{\epsilon, \delta} \implies \frac{c}{\sqrt{n}} \leq \frac{c}{\sqrt{N}} \leq \epsilon$, and $c > 0$ is not arbitrary, we can always select c such that:

$$\Phi\left(-\frac{c}{\sigma}\right) \leq \frac{\delta}{4} \quad (29)$$

$$\begin{aligned} \Pr(|\bar{X}_n - \mu| \leq \epsilon) &\geq \Pr\left(|\bar{X}_n - \mu| \leq \frac{c}{\sqrt{n}}\right) \quad \text{LHS is looser bound} \\ &\geq \Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right) - \frac{\delta}{2} \\ &\geq 1 - \frac{2\delta}{4} - \frac{\delta}{2} \quad \text{QUESTION: how may you derive this?} \\ &= 1 - \delta \end{aligned} \quad (30)$$

QUESTION: what if we change the relationship between c and δ to be $\Phi\left(-\frac{c}{\sigma}\right) \leq \frac{\delta}{3}$ instead?

6 Moment Generation Function

Introduction to MGF

6.1 Using MGF to prove Central Limit Theorem

Theorem 5 Let X_1, \dots, X_n be i.i.d R.V with $\mathbb{E}[X_k] = \mu$ and $\text{Var}(X_k) = \sigma^2 \leq \infty$, and $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, then:

$$\begin{aligned} \sqrt{n}(\bar{X} - \mu) &\xrightarrow{d} \mathcal{N}(0, \sigma^2) \\ \implies \sqrt{n} \left(\frac{\bar{X} - \mu}{\sigma} \right) &\xrightarrow{d} \mathcal{N}(0, 1) \end{aligned} \tag{34}$$

note μ and σ refer to a single R.V.

note that $X_n \xrightarrow{d} X$ means convergence in distribution, i.e.,

$$\lim_{n \rightarrow \infty} \Pr_n(X_n \leq x) = \Pr(X \leq x) \tag{35}$$

in words, as n goes to infinity, CDF of X_n converge to that of the X

6.1.1 proof

QUESTION: How do you proof convergence by distribution via MGF?

we begin, to remove notation clarity, we remove $\lim_{n \rightarrow \infty}$:

$$\begin{aligned}
\text{MGF}_{\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)}(\lambda) &= \mathbb{E}\left[\exp^{\lambda\left(\frac{\sqrt{n}}{\sigma}(\bar{X}-\mu)\right)}\right] \\
&= \text{MGF}_{(\bar{X}-\mu)}\left(\frac{\lambda\sqrt{n}}{\sigma}\right) \\
&= \text{MGF}_{\left(\frac{\sum_{i=1}^n X_i - n\mu}{n}\right)}\left(\frac{\lambda\sqrt{n}}{\sigma}\right) \\
&= \text{MGF}_{\left(\sum_{i=1}^n (X_i - \mu)\right)}\left(\frac{\lambda}{\sigma\sqrt{n}}\right) \\
&= \left(\text{MGF}_{(X_i - \mu)}\left(\frac{\lambda}{\sigma\sqrt{n}}\right)\right)^n \quad \text{property of MGF}
\end{aligned} \tag{37}$$

so we use Taylor approximation of $\text{MGF}_{(X_i - \mu)}\left(\frac{\lambda}{\sigma\sqrt{n}}\right)$ at $\lambda_0 = 0$. We need to use Taylor expansion here, as we are not after a specific moment.

QUESTION: just for fun: let's try $X_i \sim \mathcal{N}(\mu, \sigma^2)$, and MGF of general 1-D Gaussian is $\exp^{\lambda\mu + \frac{1}{2}\sigma^2\lambda^2}$

7 references

in this tutorial, I have paraphrased a number of existing courses and notes, I encourage people to see the original notes too.

1. <https://engineering.purdue.edu/ChanGroup/ECE645Notes/StudentLecture04.pdf>
2. <http://www.dklevine.com/archive/strong-law.pdf>
3. various Wikipedia pages