Machine Learning Theory: Introduction

Richard Xu

August 17, 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:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{1}$$

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

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2 \tag{2}$$

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

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

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

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

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

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

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 $\eta = \frac{1}{\beta}$:

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

QUESTION: how do you get the red part?

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle \\ f(x^*) &\geq f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle \\ \implies -f(x^*) &\leq -f(x_t) + \langle \nabla f(x_t), x_t - x^* \rangle \\ \implies f(x_t) &\leq f(x^*) + \langle \nabla f(x_t), x_t - x^* \rangle \end{aligned} \tag{6}$$

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

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

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

$$f(x_{t+1}) \leq f(x^*) + \frac{1}{2}\Big(\|x_t - x^*\|_2^2 + \|\nabla f(x_t)\|_2^2 - \|x_t - x^* - \nabla f(x_t)\|_2^2 - \eta\|\nabla f(x_t)\|_2^2\Big) \tag{8}$$

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

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

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

$$\begin{aligned} f(x_{t+1}) &\leq f(x^*) + \frac{1}{2\eta}\Big(\|x_t - x^*\|_2^2 + \cancel{\|\eta\nabla f(x_t)\|_2^2} - \|x_t - x^* - \eta\nabla f(x_t)\|_2^2 \cancel{- \|\eta\nabla f(x_t)\|_2^2}\Big) \\ &= f(x^*) + \frac{1}{2\eta}\Big(\|x_t - x^*\|_2^2 - \|x_t - x^* - \eta\nabla f(x_t)\|_2^2\Big) \\ &= f(x^*) + \frac{1}{2\eta}\Big(\|x_t - x^*\|_2^2 - \|\underbrace{x_t - \eta\nabla f(x_t)}_{x_{t+1}} - x^*\|_2^2\Big) \\ \implies f(x_{t+1}) - f(x^*) &\leq \frac{1}{2\eta}\Big(\|x_t - x^*\|_2^2 - \|\underbrace{x_t - \eta\nabla f(x_t)}_{x_{t+1}} - x^*\|_2^2\Big) \end{aligned} \tag{10}$$

finally:

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

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 $\mathcal{R}^+$:*

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

proof for this can be understood by:

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

3.1 second proof

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

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

$$\begin{aligned}
& a\mathbb{1}_{(X \geq a)} \leq X \\
\implies & \mathbb{E}\big[a\mathbb{1}_{(X \geq a)}\big] \leq \mathbb{E}[X] \\
\implies & a\mathbb{E}\big[\mathbb{1}_{(X \geq a)}\big] \leq \mathbb{E}[X] \\
\implies & a\Pr(X \geq a) \leq \mathbb{E}[X] \\
& \qquad \text{think } \mathbb{1}_A \text{ is Bernoulli R.V. with parameter } \Pr(A) \implies \mathbb{E}[1_A] = \Pr(A) \\
\implies & \mathbb{E}[X] \geq \frac{\Pr(X \geq a)}{a}
\end{aligned} \tag{16}$$

4 Chebyshev's inequality

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

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

QUESTION: what is the relationship between event $|X-\mathbb{E}(X)|\geq$ and $(X-\mathbb{E}(X))^2$?

4.1 Useful Fact

1. $\Pr(|X-\mathbb{E}(X)|\geq\epsilon)=\Pr\big(X-\mathbb{E}(X))^2\geq\epsilon^2\big)$, 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.,:

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

then you can claim:

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

4.2 alternative expressions of Chebyshev's inequality

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

4.3 application of Chebyshev's inequality

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

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

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

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

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

$$
\begin{aligned}
\Pr(|X - \mathbb{E}(X)| \geq \epsilon) &\leq \frac{\text{Var}(X)}{\epsilon^2} \\
\implies \Pr\left(\left|\bar{X}_n - \mathbb{E}(\bar{X})\right| \geq \epsilon\right) &\leq \frac{\frac{\sigma^2}{n}}{\epsilon^2} \quad \text{sub Var}(X) \to \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n\epsilon^2}
\end{aligned} \tag{23}
$$

this means that $\bar{X}_n \xrightarrow{p} \mathbb{E}(\bar{X})$, as $n \to \infty$, i.e., $\bar{X}_n$ converge in probability to $\mathbb{E}(\bar{X})$ as $n \to \infty$
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

$$X_n \xrightarrow{\text{a.s.}} X \quad \Longrightarrow \quad X_n \xrightarrow{\text{p}} X \quad \Longrightarrow X_n \xrightarrow{\text{d}} X \tag{24}$$

5.1 Convergence in probability

these are equivalent:

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

5.2 Example of $X_n \xrightarrow{\text{d}} X$: 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:

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

QUESTION: what type of convergence is above?

this is an example of $X_n \xrightarrow{\text{d}} X$: means that the limit of CDF of $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ must be equal that of Gaussian, for any interval (a, b):

5.3 Show CLT implies WLLN

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

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

think about the above: When n is small, probability for $\bar{X}_n - \mu$ to fall between $\left(\frac{-c}{\sqrt{n}}, \frac{-c}{\sqrt{n}}\right)$ **may** be less than $\Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right)$

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

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

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:

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

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

for symmetric pdf:

$$\begin{aligned}
\Phi\left(\frac{c}{\sigma}\right) + \Phi\left(-\frac{c}{\sigma}\right) &= 1 \\
\implies \Phi\left(\frac{c}{\sigma}\right) &= 1 - \Phi\left(-\frac{c}{\sigma}\right) \\
\Phi\left(\frac{c}{\sigma}\right) - \Phi\left(-\frac{c}{\sigma}\right) &= 1 - 2\Phi\left(-\frac{c}{\sigma}\right) \\
&\geq 1 - \frac{\delta}{2} \quad \text{using } \Phi\left(-\frac{c}{\sigma}\right) \leq \frac{\delta}{4}
\end{aligned} \tag{31}$$

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

5.3.1 explain the arbitrary choice of ϵ and δ

$$\forall \varepsilon, \delta \quad \exists N_{\varepsilon,\delta} \quad \text{s.t. } P(|X_n| \geq \epsilon) \leq \delta \quad \forall n > N_{\varepsilon,\delta} \tag{32}$$

$$\begin{aligned}
\delta \to 0 &\implies c \to \text{large} \qquad \text{as } \Phi\left(-\frac{c}{\sigma}\right) \leq \frac{\delta}{4} \\
\epsilon \to 0 &\implies n \to \text{large} \qquad \text{as } \frac{c}{\sqrt{n}} \leq \epsilon
\end{aligned} \tag{33}$$

note that as expected, δ is not computed as a function of ϵ. the two are obtained independently, via changing value of c, n, (the special case is not just n)

5.4 example where

6 Moment Generation Function

Introduction to MGF

6.1 Using MGF to prove Central Limit Theorem

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

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

note μ and σ refer to a single R.V.

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

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

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?
if we can prove:

$$\begin{aligned} \text{MGF}_{\lim_{n\to\infty}\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)}(\lambda) &= \text{MGF}_{X_i \sim \mathcal{N}(0,1)}(\lambda) \\ &= \exp\left(\frac{\lambda^2}{2}\right) \end{aligned} \tag{36}$$

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

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

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

$$
\begin{aligned}
&\mathbb{E}\Big[f_\lambda(0) + \lambda f'_\lambda(0) + \frac{1}{2!}\lambda^2 f''_\lambda(0) + \frac{1}{3!}\lambda^3 f'''_\lambda(0) + O(\cdot)\Big] \\
&= \mathbb{E}\Big[1 + \frac{\lambda}{\sigma\sqrt{n}}(X_i-\mu) + \frac{\lambda^2}{2\sigma^2 n}(X_i-\mu)^2 + \frac{\lambda^3}{3!\sigma^2 n^{3/2}}(X_i-\mu)^3 O(\cdot)\Big] \\
&= 1 + \frac{\lambda}{\sigma\sqrt{n}}\underbrace{\mathbb{E}\Big[(X_i-\mu)\Big]}_{=0} + \frac{\lambda^2}{2\sigma^2 n}\underbrace{\mathbb{E}\Big[(X_i-\mu)^2\Big]}_{\sigma^2} + \frac{\lambda^3}{3!\sigma^2 n^{3/2}}\mathbb{E}\Big[(X_i-\mu)^3\Big] + O(\cdot) \\
&= 1 + \frac{\lambda^2}{2n} + O(\cdot)
\end{aligned} \tag{38}
$$

taking limit and higher order terms will disappear as faster as n in denominator, and also substitute $X_i \to \bar{X}$:

$$
\begin{aligned}
&\lim_{n\to\infty} \mathrm{MGF}_{(\bar{X}-\mu)}\Big(\frac{\lambda}{\sigma\sqrt{n}}\Big) \\
&= \lim_{n\to\infty}\Big(1 + \frac{\lambda^2}{2n}\Big)^n \\
&= \exp\left(\frac{\lambda^2}{2}\right) \quad \lim_{n\to\infty}\Big(1+\frac{x}{n}\Big)^n = \exp(x)
\end{aligned} \tag{39}
$$

QUESTION: just for fun: let's try $X_i \sim \mathcal{N}(\mu, \sigma^2)$, and MGF of general 1-D Gaussian is $\exp^{\lambda\mu + \frac{1}{2}\sigma^2\lambda^2}$ using Eq.(37):

$$
\begin{aligned}
\mathrm{MGF}_{\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)}(\lambda) &= \left(\mathrm{MGF}_{(X_i-\mu)}\Big(\frac{\lambda}{\sigma\sqrt{n}}\Big)\right)^n \quad \text{property of MGF} \\
&= \left(\exp^{\frac{\sigma^2\left(\frac{\lambda}{\sigma\sqrt{n}}\right)^2}{2}}\right)^n \quad \text{using MGF of Gaussian } \exp^{\lambda\mu+\frac{1}{2}\sigma^2\lambda^2} \\
&= \left(\exp^{\left(\frac{\lambda^2}{2n}\right)}\right)^n \\
&= \exp^{\left(\frac{\lambda^2}{2}\right)}
\end{aligned} \tag{40}
$$

note that there is no need to even take $\lim_{n\to\infty}$

7 homework

Read up the following:

1. Use MGF to bound: Chernoff bound
2. hoeffding lemma
3. Bound the function value: McDiarmid's inequality
4. hoeffding inequality
5. Bernstein inequalities
6. and general concept of Rademacher Complexity

8 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