Advanced Linear Algebra Week 6 Day 1

2018/10/15 – Jonathan Hayase, updated by Prof. Weiqing Gu

1 Applications to Advanced Machine Learning

Today: Concentration Inequality

Deals with deviation of a function of independent random variables from their expectation. We start with $f : \mathbb{R} \to \mathbb{R}$ then in multi we saw $f : \mathbb{R}^n \to \mathbb{R}^m$. In this class we saw $f : \mathsf{M}_{n\times m}(\mathbb{R}) \to \mathsf{M}_{n\times m}(\mathbb{R})$. Today, we will study $f : \chi^n \to \mathbb{R}$ where Where $x_1, x_2, \ldots, x_n$ take values from χ space.

Now consider the

$$x_1 + \cdots + x_n \to f(x) = f(x_0) + (x - x_0)\nabla f(x - x_0) + \frac{1}{2!}(x - x_0)^T \nabla^2 f(x - x_0)(x - x_0) + \cdots$$

Recall the Law of Large Numbers of Probability Theory. The sum of independent random variables are, under very mild condition[1], close to their expectation with large probability.

Classically, we are interested in $\sum_{i=1}^n x_i$. Recently, $f(x_1, \ldots, x_n) = z$ where $x_1, \ldots, x_n$ are independent random variables, $f : \chi^n \to \mathbb{R}$. For example, consider the random variables forming a matrix $X = \left(x_{ij}\right)$, where x_{ij} are all independent. And $f(X) = \operatorname{tr}\left(X^T AX\right)$ for fixed A.

Let $x_1, \ldots, x_n$ be independent random variables in χ. Let $f : \chi^n \to \mathbb{R}$ and $z = f(x_1, \ldots, x_n)$.

Q: How large are "typical" deviations of Z from $\mathbb{E}Z$.

Consider $\mathbb{P}\left\{Z > \mathbb{E}[z] + t\right\}$ and $\mathbb{P}\left\{Z < \mathbb{E}[z] - t\right\}$ for $t > 0$.[2]

1.1 Markov Inequality

If $Z \geq 0$ then $\mathbb{P}\{Z \geq \mathbb{E}Z + t\} \leq \frac{\mathbb{E}Z}{t}$.

Trick: In application of you don't know $Z \geq 0$.

Because $Z \geq 0$, $\mathbb{E}Z \geq 0$, so $Z \geq \mathbb{E}Z + t \geq t$. Claim $\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}Z}{t}$ or $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{a}$.

Proof: (we use an indicator function.)

Consider a diagram plotting $F(x)$ vs a. There are two cases $x < a$ and $x \geq a$. For case 1, $a\mathbb{1}_{X\geq a} = a0 = 0 \leq x$ and $a\mathbb{1}_{x\geq a} = a1 = a \leq x$. For both cases, $x \geq a\partial_{x\geq a}$, so $\mathbb{E}[x] \geq \mathbb{E}[a\mathbb{1}_{x\geq a}] = a\mathbb{E}[\mathbb{1}_{x\geq a}]$, so $\mathbb{E} \geq a\mathbb{P}\{x \geq a\}$ and $\mathbb{P}\{x \geq a\} \leq \frac{\mathbb{E}[x]}{a}$, as desired.

[1] I don't really know what this means

[2] $\mathbb{P}$ and Pr are equivalent notation and will be used interchangably in this course.

1.1.1 Example

Give you intuition of Markov Inequality.

Suppose we have a die.

x	even $\mathbb{P}\{x = x\}$	uneven $\mathbb{P}\{x = x\}$
1	$\frac{1}{6}$	0
2	$\frac{1}{6}$	0
3	$\frac{1}{6}$	$1/2$
4	$\frac{1}{6}$	$1/2$
5	$\frac{1}{6}$	0
6	$\frac{1}{6}$	0

In the first case $\mathbb{E}[x] = 3.5$.

Interested in $\mathbb{P}(x \geq 6) = P(x = 6)$.

$$\mathbb{E}[x] = \sum x\mathbb{P}(x = x) = 1\mathbb{P}(x = 1) + \cdots + 6\mathbb{P}(x = 6) \geq 6\mathbb{P}(x = 6)$$

Suppose Markov Inequality does not hold, then

$$\mathbb{P}(x \geq 6) > \frac{3.5}{6}.$$

Then $\mathbb{E}[x] \geq 6\mathbb{P}(x = 6) > 6 \cdot \frac{3.5}{6} = 3.5$, which is a contradiction.

1.2 Chebyshev's Inequality

$$\mathbb{P}\left(|x - \mu| \geq a\right) \leq \frac{\mathrm{Var}(x)}{a^2}$$

Proof

$$\begin{aligned}
\mathbb{P}(|x - \mu| \geq a) &= \mathbb{P}\left(|x - u|^2 \geq a^2\right) \\
&\leq \frac{\mathbb{E}\|x - \mu|^2]}{a^2} && \text{by Markob inequality} \\
&= \frac{\mathbb{E}[(x - \mu)^2]}{a^2} \\
&= \frac{\mathrm{Var}(x)}{a^2}
\end{aligned}$$

1.2.1 Applications of Chebyshev's Inequality

1. Weak Law of Large Numbers

   $$\lim \mathbb{P}(|\overline{x}_N - \mu| > \epsilon) = 0$$

   Proof: Use Chebyshev's Inequality

   $$\mathbb{P}\left(|\overline{x}_N - \mu| > \epsilon\right) \leq \frac{\mathrm{Var}(\overline{x}_N)}{\epsilon^2} = \frac{\sigma^2}{N\epsilon^2} \to 0$$

   as $N \to \infty$. Where $\overline{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\mathrm{Var} = \frac{1}{N^2}\sum_{i=1}^{N} \mathrm{Var}(x_i) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$.

2. Chernoff bounds Let $x_1, \ldots, x_n$ be independent random v. real variables. By independence, we have $\mathrm{Var}(z) = \sum_{i=1}^{N} \mathrm{Var}(x_i)$. Now if they are identically distributed, then

$$\mathrm{Var}\left(\sum x_i\right) = n\,\mathrm{Var}(x_1) \qquad \text{and} \qquad \mathbb{E}\left(\sum x_i\right) = n\mathbb{E}[x_1]$$

So

$$\begin{aligned}
\mathbb{P}\left\{\left|\sum_{i=1}^{n} x_i - n\mathbb{E}[x_1]\right| \geq t\right\} &\leq \frac{\mathrm{Var}(x)}{t^2} && \text{Chebyshev's Inequality} \\
&= \frac{n\,\mathrm{Var}(x_1)}{t^2} \\
\mathbb{P}\left\{\left|\sum_{i=1}^{n} x_i - n\mathbb{E}[x_1]\right| \geq t\sqrt{n}\right\} &\leq \frac{\mathrm{Var}(x)}{\left(t\sqrt{n}\right)^2} = \frac{\mathrm{Var}(x_1)}{t^2} && \text{Chebyshev's Inequality} \\
&\leq \exp\left(-2t^2\,\mathrm{V}ar(x_1)\right) && \text{central limit theorem}
\end{aligned}$$

So we expect an exponential ail decreasing in $t^2/\mathrm{Var}(x_1)$.

Trick: Use Markov's Inequality in a more clever way. If $\lambda > 0$.

$$\begin{aligned}
\mathbb{P}(Z - \mathbb{E}Z > t) &= \mathbb{P}\left(e^{\lambda(z-\mathbb{E}Z)} > e^{\lambda t}\right) && \text{since exponential is convex} \\
&\leq \frac{\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}}{e^{\lambda t}}.
\end{aligned}$$

Now generate bounds for the moment generating function $\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}$ and optimize λ. We can show if $x_1, \ldots, x_n \in [0,1]$, then

$$\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)} \leq e^{\lambda^2/8}.$$

If $Z = \sum_{i=1}^{n} x_i$ for independent x_i, then

$$^{\lambda Z} = \mathbb{E}\prod_{i=1}^{n} e^{\lambda x_i} = \prod_{i=1}^{n} \mathbb{E}e^{\lambda x_i}$$

Now, it suffices to find $\mathbb{E}e^{\lambda x_i}$.

2 Bounded Difference Inequality

Suppose $Z_1, \ldots, Z_n$ are independent random variables taking values in some space $\mathcal{Z}$ and $f : Z^n \to \mathbb{R}$ is a function that satisfies for all i

$$\sup_{z_1,\ldots,z_n,z_i'} \left\{\left|f(z_1,\ldots,z_{i-1},z_i,z_{i+1},\ldots,z_n) - f(z_1,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n)\right|\right\} \leq x_i$$

for some constant $c_1, \ldots, c_n$. Then we have

$$\mathbb{P}\left\{\left|f(z_1^m) - \mathbb{E}[f(z_1^m)]\right| \geq t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

3 Rademacher Average

Goal want to bound the difference between empirical and true expectations uniformly over some function class G. In the context of classification or regression, we are typically interested in a class g that is the loss class associated with some function class $\mathscr{F}$.

i.e. given a bounded loss function: $\phi : D \times y \to [0,1]$ we consider the class

$$\phi_{\mathscr{F}} : \big\{(x,y) \mapsto \phi(f(x),y)\big\} \mid f \in \mathscr{F}$$

Rademacher average gives us a powerful tool to obtain uniform convergence results.

$$\mathbb{E}\left(\sup\left(\mathbb{E}[g(z)] - \frac{1}{m}\sum_{i=1}^{m} g(z_i)\right)\right)$$

where $z, \{z_i\}_{i=1}^{m}$ are i.i.d. in some space $\mathscr{F}$. Here $g \in [0,1]^2$.

By the Bounded difference inequality, the random quantity

$$\sup\left(\mathbb{E}[g(z)] - \frac{1}{m}\sum_{i=1}^{m} g(z_i)\right)$$

will be close to the above expectation with high probability.

Let $\epsilon_1, \dots, \epsilon_m$ be i.i.d. $\{\underline{t}\}$-values random variable, w/ $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$. They are also independent of the sample $z_1, \dots, z_m$.

Define the empirical Rademacher average of g as

$$\hat{\mathrm{R}}\mathrm{m}(g) \triangleq \mathbb{E}[\hat{\mathrm{R}}\mathrm{m}(g)]$$

Theorem

$$\mathbb{E}\left[\sup\left(\mathbb{E}[g(z)] - \frac{1}{m}\sum_{i=1}^{m} g(z_i)\right)\right] \le 2\,\hat{\mathrm{R}}\mathrm{m}(g)$$

4 Topic: The Johnson–Lindenstrauss Theorem

Theorem: For any $0 < \epsilon < 1$ and any integer n let k be a positive integer such that

$$k \ge 4\left(\epsilon^2/2 - \epsilon^3/3\right)^{-1} \ln n$$

Then for any set V in $\mathbb{R}^d$ of n points, there exists a linear functional (or projection) $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in V$,

$$(1-\epsilon)\|u-v\|^2 \le \big\|f(u) - f(g)\big\|^2 \le (1+\epsilon)\|u-v\|^2$$

Furthermore, this map can be found in randomized polynomial time.

Proof: if $k \ge d$ then the theorem is trivial. Suppose $k < d$. Take a random k-dimensional subspace S and we let V_i' be the projection of $V_i \in V$ into S. Then setting $L = \left\|V_i' - V_j'\right\|^2$ and

$$\mu = \frac{k}{d}\big\|V_i - V_j\big\|$$

$$\mathbb{P}[L \leq (1-\epsilon\mu)] = \mathbb{P}\left[\left\|V_i' - V_j'\right\|^2 \leq (1-\epsilon)\frac{k}{d}\left\|V_i - V_j\right\|^2\right]$$

Then by a lemma

$$\exp\left[\frac{k}{2}\left(1-(1-\epsilon)+\ln(1+\epsilon)\right)\right] \leq \exp\left(\frac{k}{2}\left(\epsilon - \left(1+\epsilon\frac{\epsilon^2}{2}\right)\right)\right) = \exp\left(\frac{k\epsilon^2}{4}\right) \leq \exp\left(-2\ln n\right) = \frac{1}{n^2}$$

$$\begin{aligned}
k &\geq 4\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)^{-1} \ln n \\
&= 4(\epsilon^2)^{-1}\left(\frac{1}{2} - \frac{\epsilon}{3}\right)^{-1} \ln n \\
\frac{k\epsilon^2}{4} &\geq \left(\frac{1}{2} - \frac{\epsilon}{3}\right)^{-1} \ln n \\
-\frac{k\epsilon^2}{4} &\leq -\left(\frac{1}{2} - \frac{\epsilon}{3}\right)^{-1} \ln n \\
&\leq \left(\frac{-1}{2}\right)^{-1} \ln n \\
&\leq -2\ln n
\end{aligned}$$

Lemma let $k \leq d$ then

1. If $\beta < 1$ then
$$\mathbb{P}\left(L \leq \frac{\beta k}{d}\right) \leq \beta^{1/2}\left(1 + \frac{(1-\beta)^2}{d-k}\right)^{\frac{d-k}{2}} \leq \exp\frac{k}{2}\left(1-\beta+\ln\beta\right)$$

2. If $\beta \geq 1$, then
$$\mathbb{P}\left(L \geq \frac{\beta k}{d}\right) \leq \beta^{1/2}\left(1 + \frac{(1-\beta)^2}{d-k}\right)^{\frac{d-k}{2}} \leq \exp\frac{k}{2}\left(1-\beta+\ln\beta\right)$$

   Proof: Markov inequality.

Setting $f(V_i) = \sqrt{\frac{d}{k}}V_i'$. Similarly

$$\mathbb{P}[L \geq (1+\epsilon)\mu] \leq 1/n^2 = \frac{\left\|V_i' - V_j'\right\| \frac{k}{d}\left\|V_i - V_j\right\|}{.}$$

By above calculations, for some fixed pair i, j the chance that the distribution $\left\|f(V_i) - f(V_j)\right\| / \left\|V_i - V_j\right\|$ does not line in the range $[1-\epsilon, 1+\epsilon]$ is at most $2/n^2$.

Using the uniform bound, the chance that some pair of point suffers a large deviation is at most $\binom{n}{s} \cdot \frac{2}{n^2} = \frac{n-1}{n} = 1 - \frac{1}{n}$.

Hence f has the desired properties w/ probability $1 - (1 - 1/n) = 1/n$.

Repeating this projection $O(n)$ times can boost the success probability to the desired constant giving us the claimed randomized polynomial time algorithm, as desired.

5 Conner's Thesis

TBD