1 Concentration of Averages

Concentration of measure is very useful in showing bounds on the errors of machine-learning algorithms. We will begin with a basic concentration inequality, which shows the concentration of measure of averages of a number of independent random variables.

**Theorem 1 (Hoeffding’s Inequality)** Let \( X_1, \ldots, X_n \) be independent and bounded random variables such that \( a_i \leq X_i \leq b_i \). Then,

\[
\Pr \left( \left| \frac{X_1 + \ldots + X_n}{n} - \mathbb{E} \left( \frac{X_1 + \ldots + X_n}{n} \right) \right| \geq \epsilon \right) \leq 2e^{-\epsilon^2 n^2 / \sum_{i=1}^{n} (b_i - a_i)^2}
\]

**Example 1: Estimating the Bias of a Coin.** Consider a coin with bias \( p \), and suppose we toss it \( n \) times. If \( X \) is the number of heads obtained, Hoeffding’s Inequality gives us:

\[
\Pr (|X - p| \geq \epsilon) \leq 2e^{-\epsilon^2 n}
\]

2 Concentration of Lipschitz Functions

Hoeffding’s Inequality shows that the mean of \( n \) independent random variables is tightly concentrated around their expectation. It turns out that similar concentration bounds can be obtained for smooth or Lipschitz functions.

**Definition 1** A function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is said to be \( \lambda \)-Lipschitz wrt to the \( L_p \)-metric if for all \( x \) and \( y \),

\[
|f(x) - f(y)| \leq \lambda \|x - y\|_p
\]

We will only consider functions which are Lipschitz with respect to the \( L_1 \) and the \( L_2 \) metrics. For example, if \( x = (x_1, \ldots, x_n) \) then the function \( f_n(x) = \frac{1}{n} (x_1 + \ldots + x_n) \) is \( \frac{1}{n} \)-Lipschitz with respect to the \( L_1 \) metric.

**Theorem 2 (Concentration of Lipschitz Functions wrt \( L_1 \)-metric)** Let \( X_1, \ldots, X_n \) be independent and bounded random variables such that \( a_i \leq X_i \leq b_i \), and let \( f \) be a function. If \( f \) is \( \lambda \)-Lipschitz with respect to the \( L_1 \) metric, then,

\[
\Pr (|f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)]| \geq \epsilon) \leq 2e^{-\epsilon^2 / \lambda^2 \sum_{i=1}^{n} (b_i - a_i)^2}
\]

Concentration bounds can be shown for functions which are \( \lambda \)-Lipschitz with respect to the \( L_2 \) metric.
Theorem 3 (Concentration of Lipschitz Function wrt $L_2$-metric) Let $S^{d-1}$ be the surface of the unit sphere in $(d-1)$ dimensions, and let $\mu$ be the uniform measure on $S^{d-1}$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be $\lambda$-Lipschitz wrt the $L_2$ metric. Then,

$$
\mu \left( f \geq \text{median}(f) + \epsilon \right) \leq 4e^{-\epsilon^2d/2\lambda^2}
$$

Example 2: Concentration of Volume on the Sphere. Let $X \sim \mu$; let $w$ be any fixed unit vector, and let $f$ be the function:

$$
f(X) = \langle X, w \rangle
$$

Then $f$ is 1-Lipschitz wrt the $L_2$ metric, because:

$$
|f(x) - f(y)| = \langle x - y, w \rangle \leq \|w\| \cdot \|x - y\| \leq \|x - y\|
$$

Observe that $\text{median}(f) = 0$ due to symmetry. Applying the theorem above on $f(X)$ and $-f(X)$, we get that for any vector $w$,

$$
\mu (|\langle w, X \rangle| > \epsilon) \leq 8e^{-2\epsilon^2d}
$$

This implies that most of the volume of a $d$-dimensional sphere is concentrated around the equator.

We will next prove Hoeffding’s Inequality, but first we need to recall a few basic probability and geometric facts.

3 Some Basic Facts

Fact 1 (Linearity of Expectation) For any two random variables $X$ and $Y$,

$$
E[X + Y] = E[X] + E[Y]
$$

Fact 2 (Variance) For a random variable $X$,

$$
\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2
$$

Fact 3 (Linearity of Variance) If $X_1, \ldots, X_n$ are $n$ independent random variables, then:

$$
\text{Var}(X_1 + \ldots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n)
$$

Fact 4 (Union Bound) For any two events $A$ and $B$,

$$
\Pr(A \cup B) = \Pr(A) + \Pr(B)
$$

Fact 5 (Jensen’s Inequality) If $f$ is a convex function, then

$$
E[f(X)] \geq f(E[X])
$$

4 Some Basic Concentration Inequalities

As an exercise, we first look at two (weaker) concentration inequalities and their proofs.

Theorem 4 (Markov’s Inequality) For any random variable $X$, and any $a \geq 0$,

$$
\Pr(|X| \geq a) \leq \frac{E[X]}{a}
$$
Proof: Observe that $|X| \geq a \cdot 1_{|X| \geq a}$. Taking expectations on both sides, we get the inequality. □

Markov’s Inequality in turn can be applied to prove stronger concentration inequalities.

**Theorem 5 (Chebyshev’s Inequality)** For any random variable $X$,

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\text{Var}(X)}{a^2}$$

Proof: Let $Z = (X - \mathbb{E}[X])^2$. Applying Markov’s Inequality to $Z$, we get:

$$\Pr(|X - \mathbb{E}[X]| \geq a) = \Pr(Z \geq a^2) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2} = \frac{\text{Var}(X)}{a^2}$$

□

Usually Chebyshev’s Inequality gives a stronger bound than Markov’s Inequality. However, Markov’s Inequality also requires less of the random variable – it only requires $\mathbb{E}[X]$ to be finite, whereas Chebyshev’s Inequality requires both $\mathbb{E}[X]$ and $\text{Var}(X)$ to be finite.

**Example 3: Symmetric Random Walks on the Line.** Consider the following stochastic process: we start at the origin, and at each time step $t$, we take a step to the left w.p. $1/2$ and to the right w.p. $1/2$. What is our position after $n$ time steps?

More formally, for each time step $t$, we define a random variable $X_t$ to represent each step of the walk as follows.

- $X_t = +1$, with probability $1/2$
- $X_t = -1$, with probability $1/2$

Since we start at the origin, the position $S_n$ after $n$ steps is defined as:

$$S_n = X_1 + X_2 + \ldots + X_n$$

Observe that using the linearity of expectation, $\mathbb{E}[S_n] = 0$, and using the linearity of variance (as the steps $X_t$ are independent), $\text{Var}(S_n) = n$. If we apply Markov’s Inequality on $|S_n|$ we get that for $c > 1$,

$$\Pr(|S_n| \geq c\sqrt{n}) \leq \frac{\mathbb{E}[|S_n|]}{c\sqrt{n}} \leq \frac{\sqrt{\mathbb{E}[S_n^2]}}{c\sqrt{n}} \leq \frac{\sqrt{\text{Var}(S_n)}}{c\sqrt{n}} \leq \frac{1}{c}$$

Applying Chebyshev’s Inequality,

$$\Pr(|S_n| \geq c\sqrt{n}) \leq \frac{\text{Var}(S_n)}{c^2 n} \leq \frac{1}{c^2}$$

Thus Chebyshev’s Inequality provides a better bound.

## 5 Proof of Hoeffding’s Inequality

In the proof of Chebyshev’s Inequality, we used Markov’s Inequality on $|X - \mathbb{E}[X]|^2$ to get a stronger bound; to prove Hoeffding’s Inequality, we will extend this idea further. To do so, we need the concept of moment generating functions.

**Definition 2** The moment generating function $\psi(t)$ of a random variable $X$ is defined as the function:

$$\psi(t) = \mathbb{E}[e^{tX}]$$
Example 4: Moment Generating Functions. Suppose $X$ is a random variable which represents the outcome of a coin toss with bias $p$. Then the moment generating function (m.g.f) of $X$ is:

$$E[e^{tX}] = pe^t + (1 - p)$$

In general if $X$ is a discrete random variable, which takes values $x_1, \ldots, x_k$ w.p. $p_1, \ldots, p_k$, then,

$$E[e^{tX}] = p_1e^{tx_1} + p_2e^{tx_2} + \ldots + p_ke^{tx_k}$$

If $X$ is a standard normal variable, then the m.g.f of $X$ is $E[e^{tX}] = e^{t^2/2}$.

In general moment generating functions may not always be defined. But if $\psi(t)$ is defined in an interval $[-\delta, \delta]$ around 0, then,

1. All moments of $X$ are finite, and
   $$E[X^k] = \frac{\partial^k \psi}{\partial t^k} \bigg|_{t=0}$$

2. If $X$ and $Y$ are two random variables such that $\psi_X(t) = \psi_Y(t)$ for all $t \in [-\delta, \delta]$, then $X$ and $Y$ have the same cumulative frequency distribution.

Fact 6 If $X$ and $Y$ are two independent random variables, then

$$E[e^{t(X+Y)}] = E[e^{tX}] \cdot E[e^{tY}]$$

Before we prove Hoeffding’s Inequality, we need one more lemma.

Lemma 1 If $X$ is a random variable such that $E[X] = 0$ and $a \leq X \leq b$, then, for any $t > 0$,

$$E[e^{tX}] \leq e^{t^2(b-a)^2/8}$$

Proof: Recall that $e^{tx}$ is a convex function of $x$. If $x = \lambda a + (1-\lambda)b$, we can use Jensen’s Inequality to write:

$$e^{tx} \leq \lambda e^{ta} + (1-\lambda)e^{tb}$$

Plugging in $\lambda = \frac{b-x}{b-a}$, we get that:

$$e^{tx} \leq \frac{b-x}{b-a} e^{ta} + \frac{x-a}{b-a} e^{tb}$$

Taking expectations on both sides and noting that $E[X] = 0$, we get:

$$E[e^{tX}] \leq \frac{be^{ta} - ae^{tb}}{b-a}$$

We can show using simple calculus that the right hand side of this equation is at most $e^{t^2(b-a)^2/8}$. □

We are now ready to prove Hoeffding’s inequality.

Theorem 6 (Hoeffding’s Inequality, restated) Let $X_1, \ldots, X_n$ be independent and bounded random variables such that $a_i \leq X_i \leq b_i$. Then,

$$\Pr \left( \left| (X_1 + \ldots + X_n) - E[X_1 + \ldots + X_n] \right| \geq \epsilon \right) \leq 2e^{-\epsilon^2 / \sum_{i=1}^{n} (b_i-a_i)^2}$$
Proof: Let $S_n = X_1 + \ldots + X_n$, and let $Y_i = X_i - \mathbb{E}[X_i]$. Then, $a_i - \mathbb{E}[X_i] \leq Y_i \leq b_i - \mathbb{E}[X_i]$. For any $t > 0$,

$$
\Pr(S_n - \mathbb{E}[S_n] \geq \epsilon) = \Pr(Y_1 + \ldots + Y_n \geq \epsilon) = \Pr(e^{t(Y_1 + \ldots + Y_n)} \geq e^{t\epsilon}) \leq \frac{\mathbb{E}[e^{t(Y_1 + \ldots + Y_n)}]}{e^{t\epsilon}}
$$

where the last step follows from applying a Markov’s Inequality. Using the independence of moment generating functions, we get that:

$$
\mathbb{E}\left[e^{t(Y_1 + \ldots + Y_n)}\right] = \mathbb{E}[e^{tY_1}] \cdot \mathbb{E}[e^{tY_2}] \cdot \mathbb{E}[e^{tY_n}] \cdot e^{-t\epsilon}
$$

Using Lemma 1, the right hand side is at most:

$$
e^{t^2(b_1-a_1)^2/8} \cdot e^{t^2(b_2-a_2)^2/8} \cdot \ldots \cdot e^{t^2(b_n-a_n)^2/8} \cdot e^{-t\epsilon}
$$

Plugging in $t = \frac{4\epsilon}{\sum_i(b_i-a_i)^2}$, this is at most $e^{-2\epsilon^2/\sum_i(b_i-a_i)^2}$, \(\square\)