Math Camp 2: Probability Theory
Sasha Rakhlin
\textbf{\(
\sigma\)-algebra}

A \(\sigma\)-algebra \(\Sigma\) over a set \(\Omega\) is a collection of subsets of \(\Omega\) that is closed under countable set operations:

1. \(\emptyset \in \Sigma\).

2. \(E \in \Sigma\) then so is the complement of \(E\).

3. If \(F\) is any countable collection of sets in \(\Sigma\), then the union of all the sets \(E\) in \(F\) is also in \(\Sigma\).
Measure

A measure $\mu$ is a function defined on a $\sigma$-algebra $\Sigma$ over a set $\Omega$ with values in $[0, \infty]$ such that

1. The empty set has measure zero: $\mu(\emptyset) = 0$

2. Countable additivity: if $E_1, E_2, E_3, \ldots$ is a countable sequence of pairwise disjoint sets in $\Sigma$,

$$
\mu \left( \bigcup_{i=1}^{\infty} E_i \right) = \sum_{i=1}^{\infty} \mu(E_i)
$$

The triple $(\Omega, \Sigma, \mu)$ is called a measure space.
**Lebesgue measure**

The *Lebesgue measure* $\lambda$ is the unique complete translation-invariant measure on a $\sigma$-algebra containing the intervals in $\mathbb{R}$ such that $\lambda([0, 1]) = 1$. 
Probability measure

*Probability measure* is a positive measure $\mu$ on the measurable space $(\Omega, \Sigma)$ such that $\mu(\Omega) = 1$.

$(\Omega, \Sigma, \mu)$ is called a *probability space*.

A *random variable* is a measurable function $X : \Omega \rightarrow \mathbb{R}$.

We can now define probability of an event

$$P(\text{event } A) = \mu \left( \{ x : I_A(x) = 1 \} \right).$$
Expectation and variance

Given a random variable $X \sim \mu$ the expectation is

$$\mathbb{E}X \equiv \int X \, d\mu.$$  

Similarly the variance of the random variable $\sigma^2(X)$ is

$$\text{var}(X) \equiv \mathbb{E}(X - \mathbb{E}X)^2.$$
Convergence

Recall that a sequence $x_n$ converges to the limit $x$

$$x_n \to x$$

if for any $\epsilon > 0$ there exists an $N$ such that $|x_n - x| < \epsilon$ for $n > N$.

We say that the sequence of random variables $X_n$ converges to $X$ in probability

$$X_n \overset{P}{\to} X$$

if

$$P (|X_n - X| \geq \epsilon) \to 0$$

for every $\epsilon > 0$. 
Convergence in probability and almost surely

Any event with probability 1 is said to happen **almost surely**. A sequence of real random variables $X_n$ converges almost surely to a random variable $X$ iff

$$P \left( \lim_{n \to \infty} X_n = X \right) = 1.$$ 

Convergence almost surely implies convergence in probability.
Law of Large Numbers. Central Limit Theorem

\textit{Weak LLN:} if $X_1, X_2, X_3, \ldots$ is an infinite sequence of i.i.d. random variables with finite variance $\sigma^2$, then

$$
\overline{X}_n = \frac{X_1 + \cdots + X_n}{n} \xrightarrow{P} \mathbb{E}X_1
$$

In other words, for any positive number $\varepsilon$, we have

$$
\lim_{n \to \infty} P \left( \left| \overline{X}_n - \mathbb{E}X_1 \right| \geq \varepsilon \right) = 0.
$$

\textit{CLT:}

$$
\lim_{n \to \infty} \Pr \left( \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq z \right) = \Phi(z)
$$

where $\Phi$ is the cdf of $N(0, 1)$. 
Useful Probability Inequalities

Jensen’s inequality: if $\phi$ is a convex function, then

$$\phi(\mathbb{E}(X)) \leq \mathbb{E}(\phi(X)).$$

For $X \geq 0$,

$$\mathbb{E}(X) = \int_0^\infty \Pr(X \geq t)dt.$$

Markov’s inequality: if $X \geq 0$, then

$$\Pr(X \geq t) \leq \frac{\mathbb{E}(X)}{t},$$

where $t \geq 0$. 
Useful Probability Inequalities

Chebyshev’s inequality (second moment): if $X$ is arbitrary random variable and $t > 0$,

$$\Pr(|X - \mathbb{E}(X)| \geq t) \leq \frac{\text{var}(X)}{t^2}.$$ 

Cauchy-Schwarz inequality: if $\mathbb{E}(X^2)$ and $\mathbb{E}(Y^2)$ are finite, then

$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}.$$
Useful Probability Inequalities

If $X$ is a sum of independent variables, then $X$ is better approximated by $\mathbb{E}(X)$ than predicted by Chebyshev’s inequality. In fact, it’s exponentially close!

Hoeffding’s inequality:

Let $X_1, ..., X_n$ be independent bounded random variables, $a_i \leq X_i \leq b_i$ for any $i \in 1...n$. Let $S_n = \sum_{i=1}^{n} X_i$, then for any $t > 0$,

$$ Pr(|S_n - \mathbb{E}(S_n)| \geq t) \leq 2 \exp \left( \frac{-2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right) $$
Remark about \( \text{sup} \)

Note that the statement

\[
\text{with prob. at least } 1 - \delta \; , \; \forall f \in \mathcal{F} , \; |\mathbb{E} f - \frac{1}{n} \sum_{i=1}^{n} f(z_i)| \leq \epsilon
\]

is different from the statement

\[
\forall f \in \mathcal{F} , \; \text{with prob. at least } 1 - \delta \; , \; |\mathbb{E} f - \frac{1}{n} \sum_{i=1}^{n} f(z_i)| \leq \epsilon.
\]

The second statement is an instance of CLT, while the first statement is more complicated to prove and only holds for some certain function classes.
Playing with Expectations

Fix a function $f$, loss $V$, and dataset $S = \{z_1, ..., z_n\}$. The empirical loss of $f$ on this data is $I_S[f] = \frac{1}{n} \sum_{i=1}^{n} V(f, z_i)$. The expected error of $f$ is $I[f] = \mathbb{E}_z V(f, z)$. What is the expected empirical error with respect to a draw of a set $S$ of size $n$?

$$\mathbb{E}_S I_S[f] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_S V(f, z_i) = \mathbb{E}_S V(f, z_1)$$