Probability review

Daniel Hsu

COMS 4772

Linearity of expectation
Let $X = (X_1, X_2, \ldots, X_d)$ be random vector with uniform distribution on unit sphere $S^{d-1} := \{ x \in \mathbb{R}^d : \|x\|_2 = 1 \}$.

Are $X_1, X_2, \ldots, X_d$ independent?

- No! But almost . . .

What is $E(X_1)$?

- If $\sigma$ is the pdf, then for any $u = (u_1, u_2, \ldots, u_d) \in S^{d-1}$,
  \[ \sigma(u_1, u_2, \ldots, u_d) = \sigma(-u_1, u_2, \ldots, u_d). \]

- So $E(X_1) = 0$.

- Similarly, $E(X_1X_2) = E(X_1X_2X_3) = \cdots = 0$.

- Also for any distinct $i_1, i_2, \ldots \in [d]$, $E(X_{i_1}X_{i_2} \cdots) = 0$.

What is $E(X_1^2)$?

- By linearity of expectation,
  \[ E\|X\|_2^2 = \sum_{i=1}^{d} E(X_i^2). \]

- But $\|X\|_2^2 = 1$ since $X$ is a random unit vector.

- So by symmetry,
  \[ E(X_1^2) = \frac{1}{d}. \]

- Nothing special about direction $(1, 0, \ldots, 0) \in S^{d-1}$.

- For any unit vector $u \in S^{d-1}$,
  \[ E(\langle u, X \rangle^2) = \frac{1}{d}. \]
Variance

- **Variance** is expected (squared) deviation of random variable from its mean:
  \[ \text{var}(X) = \mathbb{E}[ (X - \mathbb{E}(X))^2 ] \, . \]

- Another formula: \( \text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \).

- Can deduce \( (\mathbb{E}(X))^2 \leq \mathbb{E}(X^2) \) since variance is non-negative.
  - This is special case of *Jensen’s inequality*: for any convex function \( f \) and any random vector \( X \), \( f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)) \).

- Applying to random variable \( |X - \mathbb{E}(X)| \),
  \[ \mathbb{E}|X - \mathbb{E}(X)| \leq \sqrt{\text{var}(X)} =: \text{stddev}(X) \, . \]

- E.g., for uniform random unit vector \( X \), and any \( u \in S^{d-1} \),
  \[ \mathbb{E}|\langle u, X \rangle| \leq 1/\sqrt{d} \, . \]

Covariance

- If \( X \) and \( Y \) are random variables, then for any scalars \( a, b \in \mathbb{R} \),
  \[ \text{var}(aX + bY) = a^2 \text{var}(X) + 2ab \text{cov}(X, Y) + b^2 \text{var}(Y) \]
  where
  \[ \text{cov}(X, Y) = \mathbb{E}[ (X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) ] \, . \]

- If \( X \) and \( Y \) are independent, \( \text{cov}(X, Y) = 0 \), and hence
  \[ \text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) \, . \]

- Variance of the sum of *independent* random variables is the sum of the variances.
Symmetric random walk on $\mathbb{Z}$

- Stochastic process $(S_t)_{t \in \mathbb{Z}_+}$.  
  - $S_0 := 0$  
  - For $t \geq 1$,
    \[ S_t := S_{t-1} + X_t, \]
    where $\mathbb{P}(X_t = -1) = \mathbb{P}(X_t = 1) = 1/2$. Also assume  
    $\{X_t : t \in \mathbb{N}\}$ are independent. (Called Rademacher r.v.'s.)

- $S_n = \sum_{t=1}^{n} X_t$, sum of $n$ iid Rademacher r.v.'s.  
- $\text{var}(S_n) = \sum_{t=1}^{n} \text{var}(X_t) = n$.  
- So expected distance from origin is
  \[ \mathbb{E} |S_n| \leq \sqrt{\text{var}(S_n)} \leq \sqrt{n}. \]

- Note: on some realizations, can have $|S_n| = \omega(\sqrt{n})$.  
  - But how many?

---

Tail bounds
Tail bounds

- **Markov’s inequality**: for any \( t \geq 0 \),
  \[
  \mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t}.
  \]

  - Proof:
    \[
    t \cdot 1\{|X| \geq t\} \leq |X|.
    \]

  - Application to symmetric random walk:
    \[
    \mathbb{P}(|S_n| \geq c\sqrt{n}) \leq \frac{\mathbb{E}|S_n|}{c\sqrt{n}} \leq \frac{1}{c}.
    \]

Tail bounds from higher-order moments

- **Chebyshev’s inequality**: for any \( t \geq 0 \),
  \[
  \mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\text{var}(X)}{t^2}.
  \]

  - Proof: Apply Markov’s inequality to \((X - \mathbb{E}(X))^2\).

  - Application to symmetric random walk:
    \[
    \mathbb{P}(|S_n| \geq c\sqrt{n}) \leq \frac{\text{var}(S_n)}{c^2n} \leq \frac{1}{c^2}.
    \]

    (Improvement over \(1/c\) from Markov’s.)

  - Further improvements using higher-order moments.
Chernoff bounds

- Use all moments simultaneously to obtain tail bound.
- **Moment generating function (mgf):** \( M_X : \mathbb{R} \to \mathbb{R} \cup \{+\infty\} \), defined by
  \[
  M_X(\lambda) := \mathbb{E}\exp(\lambda X) = 1 + \lambda \mathbb{E}(X) + \frac{\lambda^2}{2} \mathbb{E}(X^2) + \frac{\lambda^3}{3!} \mathbb{E}(X^3) + \cdots
  \]
- If \( M_X(\lambda) \) is finite for some \( \lambda_1 < 0 \) and \( \lambda_2 > 0 \), then:
  - \( \mathbb{E}(X^p) \) is finite for all \( p \in \mathbb{N} \).
  - Graph of \( M_X \) on \([\lambda_1, \lambda_2]\) determines the distribution of \( X \).
- Often use logarithm of \( M_X \) (a.k.a. *cumulant generating function or log mgf*):
  \[
  \psi_X(\lambda) := \ln M_X(\lambda).
  \]

### Facts about log mgf

- \( \psi_X(0) = 0 \)
- \( \psi_{aX + b}(\lambda) = \psi_X(a\lambda) + b\lambda \)
- If \( X_1, X_2, \ldots, X_n \) are independent, and \( \psi_{X_i}(\lambda) \) is finite for each \( i \), then
  \[
  \psi_{\sum_{i=1}^n X_i}(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda).
  \]
- If \( \psi_X \) is finite on interval \((\lambda_1, \lambda_2)\) for some \( \lambda_1 < 0 \) and \( \lambda_2 > 0 \), then it is infinitely differentiable on the same (open) interval.
Example of (log) mgfs

- $X \sim \text{Poi}(\mu)$ (Poisson):
  \[ P(X = k) = \frac{e^{-\mu} \mu^k}{k!}, \quad k \in \mathbb{Z}_+ . \]
  - $E(X) = \mu$, $\text{var}(X) = \mu$
  - $M_X(\lambda) = \sum_{k=0}^{\infty} \frac{e^{-\mu} \mu^k}{k!} e^{\lambda k} = \ldots = e^{\mu(e^\lambda - 1)}$
  - $\psi_X(\lambda) = \mu(e^\lambda - 1)$
  - $\psi_{X-\mu}(\lambda) = \mu(e^\lambda - \lambda - 1)$
  - For $\lambda \approx 0$, $\psi_{X-\mu}(\lambda) \approx \mu \lambda^2/2$.

- $X \sim \text{N}(\mu, \sigma^2)$ (Normal)
  - $E(X) = \mu$, $\text{var}(X) = \sigma^2$
  - $M_X(\lambda) = \int e^{\lambda x} \frac{1}{\sqrt{2\pi} \sigma^2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \ldots = e^{\mu \lambda + \sigma^2 \lambda^2/2}$.
  - $\psi_{X-\mu}(\lambda) = \sigma^2 \lambda^2/2$.

Cramer-Chernoff inequality

- For any $t \in \mathbb{R}$,
  \[ P(X \geq t) \leq \exp\left(-\sup_{\lambda \geq 0} \{ t\lambda - \psi_X(\lambda) \} \right) . \]
  - Proof: apply Markov’s inequality to $\exp(\lambda X)$,
    \[ P(X \geq t) = P(\exp(\lambda X) \geq \exp(\lambda t)) \leq \frac{E \exp(\lambda X)}{\exp(\lambda t)} , \]
    and then “optimize” the choice of $\lambda \geq 0$.
  - For any $t \geq E(X)$,
    \[ P(X \geq t) \leq \exp\left(-\sup_{\lambda \in \mathbb{R}} \{ t\lambda - \psi_X(\lambda) \} \right) . \]
    - “Proof”: when $t \geq E(X)$, the optimal $\lambda$ is always $\geq 0$.
Fenchel conjugate

- Fenchel conjugate of $f: \mathbb{R} \rightarrow \mathbb{R}$:
  \[
  f^*(t) := \sup_{\lambda \in \mathbb{R}} \{ t\lambda - f(\lambda) \}.
  \]

  - E.g., $f(\lambda) = \lambda^2/2$ has $f^*(t) = t^2/2$.
  - If $f$ is bounded above by a quadratic ("strongly smooth"), then $f^*$ is bounded below by a quadratic ("strongly convex").

- Fenchel conjugate $f^*$ is max of affine functions, hence convex.
- Cramer-Chernoff inequality: For any $t \geq \mathbb{E}(X)$,
  \[
  \mathbb{P}(X \geq t) \leq \exp(-\psi^*_X(t)).
  \]

Normal tail bound

- $\mathcal{N}(\mu, \sigma^2)$ log mgf $\psi_{X-\mu}(\lambda) = \sigma^2\lambda^2/2$ has
  \[
  \psi^*_{X-\mu}(t) = t^2/(2\sigma^2).
  \]

  - $\mathbb{P}(X \geq \mu + t) \leq \exp(-t^2/(2\sigma^2)).$
  - With probability at least $1 - \delta$,
    \[
    X \leq \mu + \sqrt{2\sigma^2 \ln(1/\delta)}.
    \]
Subgaussian random variables

- Many random variables have log mgf $\psi_{X - \mathbb{E}(X)}(\lambda)$ upper-bounded by that of $N(0, \nu)$ for some $\nu > 0$, i.e.,
  \[ \psi_{X - \mathbb{E}(X)}(\lambda) \leq \nu \lambda^2 / 2. \]

  - Such random variables are called $\nu$-subgaussian (or subgaussian with variance proxy $\nu$).
  - Hence,
    \[ \psi^*_{X - \mathbb{E}(X)}(t) \geq t^2 / (2\nu). \]
  - Example: Rademacher random variable is 1-subgaussian.

- If $X_1, X_2, \ldots, X_n$ are independent, and each $X_i$ is $\nu_i$-subgaussian, then $S := \sum_{i=1}^n X_i$ is subgaussian with variance proxy $\nu := \sum_{i=1}^n \nu_i$.
  - Get tail bound for $S$ as before.

Application to symmetric random walk

- $S_n$ is subgaussian with variance proxy $n$, so
  \[ \mathbb{P}(S_n \geq t) \leq \exp(-t^2/(2n)). \]
- Using a union bound,
  \[ \mathbb{P}(|S_n| \geq c\sqrt{n}) \leq 2 \exp(-c^2/2). \]
  - Improvement over $1/c$ from Markov’s and $1/c^2$ from Chebyshev’s (except when $c$ is very small).
Hoeffding’s inequality

- Suppose \( X \) is \([0, 1]\)-valued r.v. with \( \mathbb{E}(X) = \mu \), and \( Y \) is \( \{0, 1\}\)-valued r.v. with \( \mathbb{E}(Y) = \mu \). Then

\[
\psi_{X-\mu}(\lambda) \leq \psi_{Y-\mu}(\lambda) \leq \frac{\lambda^2}{8} = \frac{1}{2} \cdot \frac{\lambda^2}{4}.
\]

- “Proof”: calculus...

- So \([a, b]\)-valued random variables are \((b-a)^2/4\)-subgaussian.
  - E.g., \([-1, +1]\)-valued random variables are 1-subgaussian.

- Tail bound for (sums of) such random variables also called Hoeffding’s inequality.

Poisson tail bound

- (Centered) \( \text{Poi}(\mu) \) log mgf \( \psi_{X-\mu}(\lambda) = \mu(e^\lambda - \lambda - 1) \) has

\[
\psi^*_{X-\mu}(t) = \mu \cdot h(t/\mu),
\]

where \( h(x) := (1 + x) \ln(1 + x) - x \).

- Interpretable approximation of \( h \):

\[
h(x) \geq \frac{x^2}{2(1 + x/3)},
\]

so

\[
\mathbb{P}(X \geq \mu + t) \leq \exp(-\mu \cdot h(t/\mu)) \leq \exp\left(-\frac{t^2}{2(\mu + t/3)} \right).
\]

- With probability at least \( 1 - \delta \),

\[
X \leq \mu + \sqrt{2\mu \ln(1/\delta)} + \ln(1/\delta)/3.
\]
Biased random walk

- Suppose $P(X_t = -1) = \frac{1-\gamma}{2}$ and $P(X_t = 1) = \frac{1+\gamma}{2}$.
  - Extreme cases: $\gamma = 1$ or $\gamma = -1$. Completely deterministic!
  - For $\gamma$ close to 1 or $-1$, should also expect better concentration around the mean.
- Similar to Bin$(n, p)$ for $p$ close to zero or one (i.e., tossing a very biased coin $n$ times).
  - Variance is small compared to maximal range.

Using variance information

- Let $X$ satisfy $X - E(X) \leq 1$ and $\text{var}(X) \leq \nu$. For any $\lambda \geq 0$,
  \[
  \psi_{X - E(X)}(\lambda) \leq \nu(e^\lambda - \lambda - 1).
  \]
  - “Proof”: exploit monotonicity of $x \mapsto (e^x - x - 1)/x^2$.
  - $\psi_{X - E(X)} \leq \psi_{\tilde{X} - E(\tilde{X})}$ on $\mathbb{R}_+$ for $\tilde{X} \sim \text{Poi}(\nu)$.
- If $X_1, X_2, \ldots, X_n$ are independent, and each $X_i - E(X_i) \leq 1$, then log mgf of $S := \sum_{i=1}^n X_i$ is bounded above by log mgf of $\text{Poi}(\mu)$ on $\mathbb{R}_+$, where $\mu := \sum_{i=1}^n \text{var}(X_i)$.
  - Get tail bound for $S$ as before; called Bennett’s inequality or Bernstein’s inequality.
Poisson approximation

- $S = \sum_{i=1}^{n} X_i$ where $X_1, X_2, \ldots, X_n$ are iid $\text{Bern}(p)$.
- Using Bennett’s inequality:

$$\mathbb{P}(S \geq np + t) \leq \exp \left( -np(1-p) \cdot h \left( \frac{t}{np(1-p)} \right) \right).$$

- Poisson heuristic: if $p = O(1/n)$, then $\text{Bin}(n, p) \approx \text{Poi}(np)$.
- $\text{Poi}(np)$ tail bound:

$$\mathbb{P}(S \geq np + t) \leq \exp \left( -np \cdot h \left( \frac{t}{np} \right) \right).$$

- So for $p = O(1/n)$, with probability at least $1 - \delta$,

$$\frac{S}{n} - p \leq O \left( \frac{\log(1/\delta)}{n} \right).$$

Why does this work?

- log mgf bounded by that of Gaussian for $\lambda$ around zero:

$$X \sim \text{Poi}(\mu) : \psi_{X-\mu}(\lambda) = \mu(e^\lambda - \lambda - 1),$$

$$X \sim \text{Bern}(p) : \psi_{X-p}(\lambda) \leq p(1-p)(e^\lambda - \lambda - 1).$$

- Another example:

$$X \sim \text{N}(0, 1) : \psi_{X^2-1}(\lambda) = -\frac{1}{2} \ln(1 - 2\lambda) - \lambda.$$  

- In above cases, there exist $\nu, c \geq 0$ such that, for all $\lambda \in [0, 1/c)$,

$$\psi_{X-E(X)}(\lambda) \leq \frac{\nu \lambda^2}{2} \cdot \frac{1}{1 - c\lambda}.$$  

- Such random variables are called $(\nu, c)$-subgamma or subgamma with variance proxy $\nu$ and scale factor $c$.
- If $(1 - c\lambda)^{-1}$ factor omitted, then called $(\nu, c)$-subexponential.
Fenchel conjugate of log mgf for subexponential

- For \((v, c)\)-subexponential random variable \(X\):
  \[
  \psi_{X - \mathbb{E}(X)}^*(t) = \sup_{\lambda \in \mathbb{R}} \left\{ t\lambda - \psi_{X - \mathbb{E}(X)}(\lambda) \right\} \geq \sup_{\lambda \in [0, 1/c]} \left\{ t\lambda - v\lambda^2/2 \right\}.
  \]

- If \(t < v/c\), then can plug-in \(\lambda := t/v\) to obtain
  \[
  \psi_{X - \mathbb{E}(X)}^*(t) \geq t^2/(2v).
  \]

- If \(t \geq v/c\), then \(t\lambda - v\lambda^2/2\) is increasing for \(\lambda \in [0, 1/c]\), so plug-in \(\lambda := 1/c\) to obtain
  \[
  \psi_{X - \mathbb{E}(X)}^*(t) \geq t/(2c).
  \]

- Conclusion:
  \[
  \psi_{X - \mathbb{E}(X)}^*(t) \geq \min \left\{ \frac{t^2}{2v}, \frac{t}{2c} \right\}.
  \]

Chi-squared distribution

- If \(X_1, X_2, \ldots, X_k\) are iid \(N(0, 1)\), then
  \(S := \sum_{i=1}^k X_i^2 \sim \chi^2(k)\) (chi-squared with \(k\) degrees-of-freedom).

- For \(\lambda \in [0, 1/2]\),
  \[
  \psi_{X_1^2 - \mathbb{E}(X_1^2)}(\lambda) = -\frac{1}{2} \ln(1 - 2\lambda) - \lambda = \frac{1}{2} \sum_{j=2}^\infty \frac{(2\lambda)^j}{j} \leq \frac{2\lambda^2}{2} \cdot \frac{1}{1 - 2\lambda},
  \]
  so \(X_1^2\) is \((2, 2)\)-subgamma; also \((4, 4)\)-subexponential.

- Consequently, \(S\) is \((4k, 4)\)-subexponential.

- Tail bound using subexponential property:
  \[
  \mathbb{P}(S - k \geq t) \leq \exp\left(-\min \left\{ \frac{t^2}{k}, \frac{t}{2} \right\} / 8\right).
  \]

- With probability at least \(1 - \delta\),
  \[
  S \leq k + \max \left\{ \sqrt{8k \ln(1/\delta)}, 8\ln(1/\delta) \right\}.
  \]

- A tighter analysis gets a bound of \(k + 2\sqrt{k \ln(1/\delta)} + 2\ln(1/\delta)\).
Subgaussian moments

Suppose $X$ is $\nu$-subgaussian and $\mathbb{E}(X) = 0$.

- For any $k \in \mathbb{N}$,
  $$\mathbb{E} |X|^k \leq (2\nu)^{k/2} k \Gamma(k/2).$$

- **Proof:**
  \[
  \mathbb{E} |X|^k = \int_0^\infty \mathbb{P}(|X|^k \geq t) \, dt \leq \int_0^\infty 2e^{-t^2/(2\nu)} \, dt \ldots
  \]

- $X^2$ is $(128\nu^2, 8\nu)$-subexponential.

- **Proof:** Use Taylor series to express $\psi_{X^2 - \mathbb{E}(X^2)}$ in terms of even moments of $X$. 

27