Probability Review/New Material

High Dim Probability & Linear Algebra for ML and Sig Proc

Namrata Vaswani

Iowa State University

Reading material, Relevant Courses I

1. The book "High Dimensional Probability for Data Science" by Roman Vershynin; and early parts
2. The tutorial article on "Non-asymptotic Random Matrix Theory" also by Vershynin
3. Probability:
   https://www.ece.iastate.edu/~namrata/EE527_Spring14/Probability_recap_3.pdf

   Good courses to take at ISU: EE 523, STAT 542, 543, EE/Math 623X
4. Linear Algebra (based on first few chapters of Horn and Johnson, Matrix Analysis):
   https://www.ece.iastate.edu/~namrata/EE527_Spring14/linearAlgebraNotes.pdf

   Good courses to take at ISU: MATH 510 (first half of the course); if too advanced, then first take MATH 407/507.
5. Review of Basics:
   1. Probability:
      https://www.ece.iastate.edu/~namrata/EE527_Spring12/322_recap.pdf
      http://cs229.stanford.edu/section/cs229-prob.pdf
      http://cs229.stanford.edu/section/more_on_gaussians.pdf

Reading material, Relevant Courses II

2. Linear Algebra: Andrew Ng's review from CS229 course at Stanford:
   http://cs229.stanford.edu/section/cs229-linalg.pdf
   also
   http://cs229.stanford.edu/livenotes2020spring/linearalgebra-slides.pdf

Reading I

Chapter 1 of book (Vershynin's book)

Notation I

- Order etc
  - Order notation: $f(n) \in O(g(n))$ means that there exists an $n_0 < \infty$ such that for all $n > n_0$, $f(n) \leq Cg(n)$ for a numerical constant C
  - Omega notation: $f(n) \in \Omega(g(n))$ means that there exists an $n_0 < \infty$ such that for all $n > n_0$, $f(n) \geq Cg(n)$ for a numerical constant C
  - $a \ll b$ means a/b is less than $O(1)$
  - Re-use of letter C: C is used to denote different numerical constants in different uses
- Linear algebra
  - For a matrix A, A' or A^T or $A^\top$ denotes matrix transpose; other use of MATLAB notation too.
  - Sphere in $\Re^n$: $\mathcal{S}^{n-1}$, e.g., circle is a sphere in $\Re^2$ and is denoted by $\mathcal{S}^1$
  - Norms: $\|.\|$: l2-norm, $\|.\|_1$: l1-norm, $\|.\|_F$: Frobenius norm
  - Indicator function: $\mathbb{1}_{statement} = 1$ if *statement* is true and $= 0$ otherwise.
- Probability
  - For a set A, A^c denotes its complement set.
  - Cumulative Distribution Function (CDF): $F_X(x) := \Pr(X \leq x$
  - MGF $M_X(t) = \mathbb{E}[e^{tX}]$ for a scalar X. For a vector, $\underline{X}$, $M_{\underline{X}}(\underline{u}) = \mathbb{E}[e^{\underline{u}'\underline{X}}]$
  - Characteristic function: $C_X(t) = \mathbb{E}[e^{itX}]$: it is the FT of the distribution of X computed at frequency $-t$.

Notation II

- $\Pr(A, B) = \Pr(A \text{ and } B) = \Pr(A \cap B)$.
- Gaussian $\mathcal{N}(\mu, \Sigma)$
- Bernoulli with probability of a 1 p: $Bern(p)$
- Symmetric Bernoulli $SymBern$: $X = -1$ w.p. $1/2$ and $X = +1$ w.p. $1/2$
- w.h.p. :
- w.p. :

Basics: Simple algebra bounds: move to the end I

1. https://www.lkozma.net/inequalities_cheat_sheet/ineq.pdf
2. Simple algebra bounds
   - For any $x \geq 0$, $1 + x \leq e^x$
     used very often to convert $\Pi_i(1 + \mu_i)$ to $e^{\sum_i \mu_i}$ (appears when bounding MGF of sums of indep r.v.s)
   - For $0 < x < 1$, $\log(1 + x) \geq x/(1 + x/2)$
   - For all $x > -1$, $\log(1 + x) \leq x$
   - For any $x \geq 0$, $e^x < x + e^{x^2}$
     used in subGaussian properties' equivalence.
   - ?? For any $x \geq 0$, $\frac{1}{1-x} \leq e^{2x}$
   - For any $z > 0$, $\max(|z - 1|, |z - 1|^2) \leq |z^2 - 1|$
   - For any $z > 0$, $|z - 1| \geq \delta$ implies $|z^2 - 1| \geq \max(\delta, \delta^2)$.
   - Stirling /factorial bounds
     - ★ $\Gamma(x) < x^x$ $\Gamma(x) :=??$
     - ★ $p! > (p/e)^p$, easy to see that $p! < p^p$
     - ★ $(\frac{n}{k})^k \leq \binom{n}{k} \leq \sum_{k'=0}^{k} \binom{n}{k} \leq (\frac{en}{k})^k$
   - Taylor series
     - ★ $\exp(x) = \sum_{p=0}^{\infty} \frac{x^p}{p!}$
3. Copy more from page 23, 30 of Vershynin book. TBD

Basics: Probability concepts assumed I

1. Probability concepts assumed:
   Probability axioms, disjoint events, independent events, conditional probability define, DeMorgan's laws, counting arguments

   Use: try to convert an exact probability computation into probability of union of disjoint events, or intersection of independent events, or some combination of these ideas.

   For upper bounding $\Pr(\cup_i A_i)$: use union bound

   For lower bounding $\Pr(\cup_i A_i)$: use DeMorgan's + independence, and lower bounds on $\Pr(A_i)$ or
   use $\Pr(A) \geq \Pr(A, B)$ followed by lower bound $\Pr(B)$ and $\Pr(A|B)$ (see use of this in the random vectors' theorem).

   Many more ideas of course

   Random variables: define PMF, joint PMF, PDF, joint PDF, CDF, joint CDF. Conditional CDF, conditional PDF.

Quick test of concepts: Given random variables (r.v.) $X_1, X_2, \ldots X_n$.

1. Compute distribution of $Z = |X_1 + 1|$
2. Compute distribution of $Z = X_1 mod 5$ (remainder when X_1 is divided by 5.
3. Compute the distribution of $Z = X_1 + X_2$. First
4. Compute the distribution of the smallest r.v., $Z = \min(X_1, X_2, ...X_n)$.
5. Compute the distribution of the second smallest r.v. (2nd order statistic).

1. Chain rule: extension of $P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$

$$\Pr(A_1, A_2, \ldots A_n) = \Pr(A_1)\Pr(A_2|A_1)\ldots\Pr(A_k|A_1, A_2, ...A_k-1)\ldots\Pr(A_n|A_1, A_2, ...A_{n-1})$$

2. Total expectation theorem for events, Law of iterated expectations for r.v.s
Consider events $A_1, A_2, \ldots, A_n$ that form a partition of the sample space. Partition means: all the events are disjoint and their union forms the entire sample space. Simplest example of a partition is $n = 2$, $A_1 = A$, $A_2 = A^c$.
We have

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X|A_i]\Pr(A_i)$$

If we set $X = \mathbb{1}_E$ for an event E,, the above gives the total probability result.

$$\Pr(E) = \sum_i \Pr(E|A_i)\Pr(A_i)$$

For two r.v.s X, Y (scalar or vector r.vs),

$$\mathbb{E}[g(X, Y)] = \mathbb{E}[\mathbb{E}[g(X, Y)|X]]$$

(here $\mathbb{E}[.]$ takes expectation w.r.t. all r.v.s - here X, Y; $\mathbb{E}[.|X]$ takes expectation conditioned on X.

Probability Review II

3. Independence and Conditional independence of events, r.v.s:
   1. Two events independent: $\Pr(A, B) = \Pr(A)\Pr(B)$
   2. A set of n events is independent if for any subset $S \subseteq [1, 2, ..n]$,

$$\Pr(\cap_{i \in S} A_i) = \prod_{i \in S} \Pr(A_i)$$

   3. A set of n r.v.s, $X_1, X_2, ...X_n$ independent iff joint distribution is equal to product of marginals

$$F_{X_1, X_2, ... X_n}(x_1, x_2, ...x_n) = \prod_{i=1}^{n} F_{X_i}(x_i)$$

   4. Conditional independence given $Z = z$ for all $z \in C$: above holds conditioned on $Z = z$ for all $z \in C$.
   5. i.i.d. : independent and $F_{X_i}(x) = F_{X_1}(x)$, so that

$$F_{X_1, X_2, ... X_n}(x_1, x_2, ...x_n) = \prod_{i=1}^{n} F_{X_i}(x_i) = \prod_{i=1}^{n} F_{X_1}(x_i)$$

6. X, Y (scalars or vectors) independent implies
$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$$
7. Conditionally independent given event C: above holds given event C. Same for conditional indep given a r.v.
8. X indep of $Y, Z \Rightarrow X$ indep Y; and X conditionally indep Y given Z.

4. Cauchy-Schwarz inequality
   1. For two vectors v_1, v_2,
   $$(v_1' v_2)^2 \leq \|v_1\|^2 \ \|v_2\|^2$$
   2. For two scalar r.v.s X, Y,
   $$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$
   3. obvious extensions apply for random vectors and matrices.
5. Union bound: for a set of events A_i, suppose that $\Pr(A_i) \geq 1 - p_i$. Then
$$\Pr(A_1, A_2, \ldots, A_n) \equiv \Pr(\cap_i A_i) = 1 - \Pr(\cup_i A_i^c) \geq 1 - \sum_i P(A_i^c) \geq 1 - \sum_i p_i$$

Probability Review IV

6. Moment Generating Function (MGF)
   $M_X(t) := \mathbb{E}[e^{tX}]$ for a scalar X.
   For a vector, $\underline{X}$, $M_{\underline{X}}(\underline{u}) = \mathbb{E}[e^{\underline{u}^T \underline{X}}]$
7. Characteristic function : $C_X(t) := \mathbb{E}[e^{itX}]$:
   it is the FT of the distribution of X computed at frequency $-t$.

Scalar Gaussian r.v.

First note that a scalar Gaussian r.v. X with mean μ and variance σ^2 has the following pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Its characteristic function can be computed by computing the Fourier transform at $-t$ to get

$$C_X(t) = e^{j\mu t} e^{-\frac{\sigma^2 t^2}{2}}$$

Gaussian random vector (Jointly Gaussian r.v.s)

Any of the following can be used as a definition of j G.

All vectors should ideally be replaced by $\underline{X}$ etc.

Gaussian r.v.s II

1. The $n \times 1$ random vector X is jointly Gaussian if and only if the scalar

$$u^T X$$

is Gaussian distributed for all $n \times 1$ vectors u

2. The random vector X is jointly Gaussian if and only if its characteristic function, $C_X(u) := \mathbb{E}[e^{iu^T X}]$ can be written as

$$C_X(u) = e^{iu^T \mu} e^{-u^T \Sigma u/2}$$

where $\mu = \mathbb{E}[X]$ and $\Sigma = cov(X)$.

- Proof idea - one side: Given X has above $C_X(u)$, show that $V := u^T X$ is G for any vector u. To do this, show that $C_V(t)$ has the G c.f. expression. To show this, use the fact that $C_V(t) = C_X(tu)$ for scalar t.
- Proof idea - other side: Given $u'X$ is G for any u. Let $V := u'X$. Its mean and variance are $\mu = u^T \mu$ and $\sigma^2 = u^T \Sigma u$ and thus $C_V(t) = e^{j\mu t} e^{-\frac{\sigma^2 t^2}{2}}$. Now, $C_X(u) = C_V(1) = e^{j\mu} e^{-\frac{\sigma^2}{2}}$. Substituting for μ, σ^2 gives the $C_X(u)$ expression we want to get.

3. The random vector X is j G if and only if it can be written as an affine function of i.i.d. standard Gaussian r.v's.
   - Proof uses $C_X(u)$ expression definition.
   - Proof: suppose $X = AZ + a$ where $Z \sim \mathcal{N}(0, I)$; get an expression for its c.f. by using the c.f. definition and the fact that Z is a vector of i.i.d. standard Gaussian scalar r.v.s and thus $\mathbb{E}[e^{itZ_j}] = e^{t^2/2}$ for any t. Show that the c.f. of X satisfies the $C_X(u)$ formula given in 2 with $\mu_X = a, \Sigma_X = AA^T$.
   - Proof (other side): suppose X is j G with mean μ_X and covariance Σ_X; X can always be expressed as $X = \Sigma^{1/2} Z + \mu$ where $Z := \Sigma^{-1/2}(X - \mu)$; show that Z is std. G (by getting an expression for its c.f.).
     (c.f. of a std G Z is $C_Z(u) = e^{||u||^2/2}$).
4. The random vector X is j G if and only if it can be written as an affine function of jointly Gaussian r.v's.
   - Proof: Suppose X is an affine function of a j G r.v. Y, i.e. $X = BY + b$. Since Y is j G, by 3, it can be written as $Y = AZ + a$ where $Z \sim \mathcal{N}(0, I)$ (i.i.d. standard Gaussian). Thus, $X = BAZ + (Ba + b)$, i.e. it is an affine function of Z, and thus, by 3, X is j G.
   - Proof (other side): X is j G. So by 3, it can be written as $X = BZ + b$. But $Z \sim \mathcal{N}(0, I)$ i.e. Z is a j G r.v.

5. The random vector X is jointly Gaussian if and only if its joint pdf can be written as

$$f_X(x) = \frac{1}{(\sqrt{2\pi})^n det(\Sigma)} e^{-(X-\mu)^T \Sigma^{-1} (X-\mu)/2} \tag{1}$$

   - Proof: follows by computing the characteristic function from the pdf and vice versa. Suppose X has above PDF. Then $C_X(u) = \mathbb{E}[\exp(iu'X)] = \int_x \exp(i\sum_j u_j x_j) f_X(x) dx$, here x is a vector. Change of variables: let $z = \Sigma^{-1/2}(x-\mu)$ and substitute into the integral. Integral will decouple into a product with term in the product being c.f. of a scalar Gaussian. Use formula, to finally get the vector Gaussian c.f. expression. Thus X is j G. Suppose X is j G. Then it has the given c.f. By uniqueness of Fourier transform, its density is given by (1).

Properties

1. If X_1, X_2 are j G, then the conditional distribution of X_1 given X_2 is also j G
2. If the elements of a j G r.v. X are pairwise uncorrelated (i.e. non-diagonal elements of their covariance matrix are zero), then they are also mutually independent.
3. Any subset of X is also j G.

1. Integral identity
   For a scalar r.v. Z that is non-negative, i.e., $Z \geq 0$ w.p. 1,

$$\mathbb{E}[Z] = \int_{\tau=0}^{\infty} \Pr(Z > \tau) d\tau$$

   Proof: Use $x = \int_{t=0}^{x} 1 dt = \int_{t=0}^{\infty} \mathbb{1}(t \leq x) dt$ followed by moving expectation inside integral sign (allowed since indicator func is bounded).
2. Use integral identity to convert w.h.p. bound to bound on expectation:
   Given a non-negative r.v. Z that satisfies $\Pr(Z > u_0 + t) \leq e^{-t^2}$ for all $t \geq 0$ ($Z \leq 1.1 u_0$ w.h.p.) for a $u_0 \gg 2$ (u_0 is more than order 1). This implies

$$\mathbb{E}[Z] \leq u_0 + 2 \leq 1.1 u_0 \text{ the second bound assumes } u_0 \gg 2$$

   - To use this second bound of $1.1u_0$, scale Z so that $u_0 \gg 2$.
   - Similarly, if we are told that $\Pr(Z < u_0 - t) \leq e^{-t^2}$ for all $t \geq 0$ ($Z \geq 0.9u_0$ w.h.p.), assuming $2 \ll u_0$, we can show that

$$\mathbb{E}[Z] \geq u_0 - 2 \geq 0.9 u_0$$

- Proof idea: apply integral identity, split integral into 0 to u_0 and then u_0 to ∞. In the first one, bound the probability by 1, in the second one, use the assumption, to get $\mathbb{E}[Z] \leq u_0 + \int_{t=0}^{\infty} e^{-t^2} dt \leq u_0 + \sqrt{2\pi}/2 < u_0 + 2$.
  Proof idea for lower bound:
  $\mathbb{E}[Z] \geq \int_{\tau=0}^{u_0} \Pr(Z > \tau) d\tau = \int_{t=0}^{u_0} \Pr(Z > u_0 - t) dt \geq \int_{t=0}^{u_0} (1 - e^{-t^2}) dt$

3. Gaussian tail bounds: $X \sim \mathcal{N}(0,1)$:

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \Pr(X \geq t) \leq \frac{1}{\sqrt{2\pi}} \frac{1}{t} e^{-t^2/2}$$

Proof idea:

- Upper bound: use $\int_{x=t}^{\infty} e^{-x^2/2} dx \leq \int_{x=t}^{\infty} (x/t) e^{-x^2/2} dx$ and then use change of variables to solve the integral.
- Lower bound: see page 12 of Vershynin book

Markov inequality and applications I

For a non-negative r.v. Z,

$$\Pr(Z > s) \leq \frac{\mathbb{E}[Z]}{s}$$

Proof: easy application of integral identity

$$\mathbb{E}[Z] \geq \int_0^s \Pr(Z > \tau) d\tau \geq \Pr(Z > s)(\int_0^s d\tau) = \Pr(Z > s)s$$

Applications: basic ideas

1. Apply this to $Z = |X - \mu|$ with $\mu = \mathbb{E}[X]$, to get Chebyshev's inequality.
2. Apply this to $Z = e^{tX}$ for any $t \geq 0$. notice e^{tX} is always non-negative.

   $$\Pr(X > s) = \Pr(e^{tX} > e^{ts}) \leq e^{-ts}\mathbb{E}[e^{tX}] = e^{-ts} M_X(t)$$

   Since this bound holds for all $t \geq 0$, we can take a $\min_{t \geq 0}$ of the RHS or we can substitute in any convenient value of t.
3. To get a bound for $\Pr(X < -s)$, use $Z = e^{-tX}$ for $t \geq 0$.

4. Useful for sums of independent r.v.s: if $S = \sum_i X_i$ with X_i's independent, then $M_X(t) = \prod_i M_{X_i}(t)$. So then we get

$$\Pr(\sum_i X_i > s) \leq \min_{t \geq 0} e^{-ts} M_{\sum_i X_i}(t) = \min_{t \geq 0} e^{-ts} \prod_i \mathbb{E}[e^{tX_i}]$$

5. Use exact expression for MGF or a bound on MGFs (e.g. Hoeffding's lemma bounds the MGF of any bounded r.v.)
6. Followed by often using $1 + x \leq e^x$ to simplify things
7. Final step: either minimizer over $t \geq 0$ by differentiating the expression or a pick a convenient value of t to substitute.
8. *disregard this in first read:* Final final step that is used sometimes: suppose get a bound $g(s)$ but want to show $g(s) \leq f(s)$ for some simpler expression $f(s)$: try to show that $g(s) - f(s)$ is a decreasing function for the desired range of s values with $g(0) - f(0) = 0$ or something similar: this is used in Chernoff inequality for $Bern(p_i)$ r.v.s. for small s setting.

Quick test of concepts: Given random variables (r.v.) $X_1, X_2, \ldots X_n$.

1. Compute distribution of $Z = |X_1 + 1|$
2. Compute distribution of $Z = X_1 mod 5$ (remainder when X_1 is divided by 5.
3. Compute the distribution of $Z = X_1 + X_2$. First
4. Compute the distribution of the smallest r.v., $Z = \min(X_1, X_2, ...X_n)$.
5. Compute the distribution of the second smallest r.v. (2nd order statistic).

Some Topics

1. Chain Rule: extension of $P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$

$$P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \ldots P(A_n|A_1, A_2, \ldots A_{n-1})$$

2. Total probability: if $B_1, B_2, \ldots B_n$ form a *partition* of the sample space, then

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Partition: The events are mutually disjoint and their union is equal to the sample space.

3. Union bound: suppose $P(A_i) \geq 1 - p_i$ for small probabilities p_i, then

$$P(\cap_i A_i) = 1 - P(\cup_i A_i^c) \geq 1 - \sum_i P(A_i^c) \geq 1 - \sum_i p_i$$

4. Independence and Conditional Independence
   - events A, B are independent iff

$$P(A,B) = P(A)P(B)$$

   - events $A_1, A_2, \dots A_n$ are mutually independent iff for any subset $S \subseteq \{1, 2, \dots, n\}$,

$$P(\cap_{i \in S} A_i) = \prod_{i \in S} P(A_i)$$

   - analogous definition for random variables: for mutually independent r.v.'s the joint pdf of any subset of r.v.'s is equal to the product of the marginal pdf's.
   - events A, B are conditionally independent given an event C iff

$$P(A,B|C) = P(A|C)P(B|C)$$

   - extend to a set of events as above
   - extend to r.v.'s as above
5. Aside: Given X is independent of $\{Y, Z\}$. Then,
   - X is independent of Y; X is independent of Z

Old recap document III

- X is conditionally independent of Y given Z
- $\mathbb{E}[XY|Z] = \mathbb{E}[X|Z]\mathbb{E}[Y|Z]$
- $\mathbb{E}[XY|Z] = \mathbb{E}[X]\mathbb{E}[Y|Z]$

6. Law of Iterated Expectations:

$$\mathbb{E}_{X,Y}[g(X,Y)] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[g(X,Y)|Y]]$$

7. Conditional Variance Identity:

$$Var_{X,Y}[g(X,Y)] = \mathbb{E}_Y[Var_{X|Y}[g(X,Y)|Y]] + Var_Y[\mathbb{E}_{X|Y}[g(X,Y)|Y]]$$

8. Cauchy-Schwartz Inequality:
   1. For vectors v_1, v_2, $(v_1'v_2)^2 \leq \|v_1\|_2^2\|v_2\|_2^2$
   2. For vectors:

$$(\frac{1}{n}\sum_{i=1}^{n} x_i'y_i)^2 \leq (\frac{1}{n}\sum_{i=1}^{n}\|x_i\|_2^2)(\frac{1}{n}\sum_{i=1}^{n}\|y_i\|_2^2)$$

   3. For matrices:

$$\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{X}_i\mathcal{Y}_i'\|^2 \leq \|\frac{1}{n}\sum_{i=1}^{n}\mathcal{X}_i\mathcal{X}_i'\|_2\|\frac{1}{n}\sum_{i=1}^{n}\mathcal{Y}\mathcal{Y}'\|_2$$

4. For scalar r.v.'s X, Y: $(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$
5. For random vectors X, Y,

$$(\mathbb{E}[X'Y])^2 \leq \mathbb{E}[\|X\|_2^2]\mathbb{E}[\|Y\|_2^2]$$

6. Proof follows by using the fact that $\mathbb{E}[(X - \alpha Y)^2] \geq 0$. Get a quadratic equation in α and use the condition to ensure that this is non-negative
7. For random matrices $\mathcal{X}, \mathcal{Y}$,

$$\|\mathbb{E}[\mathcal{X}\mathcal{Y}']\|_2^2 \leq \lambda_{\max}(\mathbb{E}[\mathcal{X}\mathcal{X}'])\lambda_{\max}(\mathbb{E}[\mathcal{Y}\mathcal{Y}']) = \|\mathbb{E}[\mathcal{X}\mathcal{X}']\|_2\|\mathbb{E}[\mathcal{Y}\mathcal{Y}']\|_2$$

   Recall that for a positive semi-definite matrix M, $\|M\|_2 = \lambda_{\max}(M)$.
8. Proof: use the following definition of $\|M\|_2$: $\|M\|_2 = \max_{x,y:\|x\|_2=1,\|y\|_2=1} |x'My|$, and then apply C-S for random vectors.
9. Hoeffding's lemma: bounds the MGF of a *zero mean* and *bounded* r.v..
   - Suppose $\mathbb{E}[X] = 0$ and $P(X \in [a, b]) = 1$, then

$$M_X(s) := \mathbb{E}[e^{sX}] \leq e^{\frac{s^2(b-a)^2}{8}} \text{ if } s > 0$$

   Proof: use Jensen's inequality followed by mean value theorem, see `http://www.cs.berkeley.edu/~jduchi/projects/probability_bounds.pdf`

Old recap document V

10. Convergence in probability. A sequence of random variables, $X_1, X_2, \ldots X_n$ converges to a constant a in probability means that for every $\epsilon > 0$,

$$\lim_{n \to \infty} Pr(|X_n - a| > \epsilon) = 0$$

11. Convergence in distribution. A sequence of random variables, $X_1, X_2, \ldots X_n$ converges to random variable Z in distribution means that

$$\lim_{n \to \infty} F_{X_n}(x) = F_Z(x), \text{ for almost all points} x$$

12. Convergence in probability implies convergence in distribution
13. Consistent Estimator. An estimator for θ based on n random variables, $\hat{\theta}_n(\underline{X})$, is consistent if it converges to θ in probability for large n.
14. Weak Law of Large Numbers (WLLN) for i.i.d. scalar random variables, $X_1, X_2, \ldots X_n$, with finite mean μ. Define

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$$

For any $\epsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

Proof: use Chebyshev if σ^2 is finite. Else use characteristic function