Probability Inequalities

Thm 1 (Gaussian Tail Inequality): Let $X \sim \mathcal{N}(0,1)$. Then

$$\mathbb{P}(|X| > \epsilon) \leq \frac{2}{\epsilon} e^{-\epsilon^2/2} \tag{1}$$

Additionally:

$$\mathbb{P}(|\overline{X}_n| > \epsilon) \leq \frac{1}{\sqrt{n}\epsilon} e^{-n\epsilon^2/2} \tag{2}$$

Thm 2 (Markov Inequality): Let X be a non-negative random variable s.t. $\mathbb{E}(X)$ exists. Then $\forall\ t > 0$

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t} \tag{3}$$

Thm 3 (Chebyshev's Inequality): Let $\mu = \mathbb{E}(X)$ and $\sigma^2 = \mathrm{Var}(X)$. Then

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \tag{4}$$

$$\mathbb{P}(|(X-\mu)/\sigma| \geq t) \leq \frac{1}{t^2} \tag{5}$$

Lemma 4: Let $\mathbb{E}(X) = 0$ and $a \leq X \leq b$. Then

$$\mathbb{E}(e^{tX}) \leq e^{t^2(b-a)^2/8} \tag{6}$$

Lemma 5: Let X be any random variable. Then

$$\mathbb{P}(X > \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon} \mathbb{E}(e^{tX}) \tag{7}$$

Thm 6 (Hoeffding's Inequality): $X_1, \ldots, X_n$ iid, $\mathbb{E}(X_i) = \mu$, $a \leq X_i \leq b$. Then $\forall \epsilon > 0$

$$\mathbb{P}(|\overline{X} - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2} \tag{8}$$

Thm 9 (McDiarmid):
Thm 12 (Cauchy-Schwartz inequality):
Thm 13 (Jensen's inequality):
Ex 15 (Kullback Leibler distance):
Thm 18?:
O_p **and** o_p**:**

Shattering

Note: remember uniform bounds and union bound.
$\mathcal{A}$ **picks out** $G \subset F$.
$S(\mathcal{A}, F)$.
F **shattered** by $\mathcal{A}$ if $S(\mathcal{A}, F) = 2^{|F|}$ (ie if $\mathcal{A}$ picks out all $G \subset F$).
The **shatter coefficient** $s_n(\mathcal{A}) = \sup_{F \in \mathcal{F}_n} s(\mathcal{A}, F)$. Note $n = |F|$ and $s_n(\mathcal{A}) \leq 2^n$.
Thm 5:
The **VC dimension** $d(\mathcal{A})$ = largest n s.t. $s_n(\mathcal{A}) = 2^n$.

Random Samples

For $X_1, \ldots, X_n \sim F$ a **statistic** is any $T = g(X_1, \ldots, X_n)$.
E.g. $\overline{X}_n$, $S_n = \sum_i (X_i - \overline{X}_n)^2/(n-1)$, $\big(X_{(1)}, \ldots, X_{(n)}\big)$
Note: $\mathbb{E}(\overline{X}_n) = \mathbb{E}(X_i)$, $\mathrm{Var}(\overline{X}_n) = \mathrm{Var}(X_i)/n$, $\mathbb{E}(S_n)^2 = \mathrm{Var}(X_i)$.
Note: sum of bernouilli is binomial(n,p), sum of exp(beta) is gamma(n,beta), sum of standard normal is chi-squared(n-dof).
Thm. 1: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2) \implies \overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$.

Convergence

$X, X_1, X_2, \ldots$ random variables.
(1) X_n converges **almost surely** $X_n \xrightarrow{a.s.} X$ if $\forall \epsilon > 0$

$$\mathbb{P}(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1 \tag{9}$$

(2) X_n converges **in probability** $X_n \xrightarrow{p} X$ if $\forall \epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0 \tag{10}$$

(3) X_n converges **in quadratic mean** $X_n \xrightarrow{qm} X$ if

$$\lim_{n\to\infty} \mathbb{E}[(X_n - X)^2] = 0 \tag{11}$$

(4) X_n converges **in distribution** $X_n \rightsquigarrow X$ if

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t) \tag{12}$$

$\forall t$ on which F_X is continuous.

Thm 7: Conv. a.s. and in q.m. imply conv. in prob. All three imply conv. in distribution. Conv. in distribution to a point-mass also implies conv. in prob.
Ex from class: Showed conv. in prob $\not\Longrightarrow$ conv. a.s.. Showed conv. in prob $\not\Longrightarrow$ conv. in q.m.. Showed conv. in distro $\not\Longrightarrow$ conv. in prob.
Thm 10a: X,X_n,Y,Y_n random variables. Then

$$(a)\ \ X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \implies X_n + Y_n \xrightarrow{p} X + Y \tag{13}$$

$$(b)\ \ X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \implies X_n Y_n \xrightarrow{p} XY \tag{14}$$

$$(c)\ \ X_n \xrightarrow{qm} X, Y_n \xrightarrow{qm} Y \implies X_n + Y_n \xrightarrow{qm} X + Y \tag{15}$$

Thm 10b (Slutzky's Thm): X,X_n,Y_n random variables. Then

$$(a)\ \ X_n \rightsquigarrow X, Y_n \rightsquigarrow c \implies X_n + Y_n \rightsquigarrow X + c \tag{16}$$

$$(b)\ \ X_n \rightsquigarrow X, Y_n \rightsquigarrow c \implies X_n Y_n \rightsquigarrow cX \tag{17}$$

Thm 12 (Law of Large Numbers): $X_1, \ldots, X_n$ iid, $\mathbb{E}(X_i) = \mu$ $\implies \overline{X}_n \xrightarrow{\text{qm}} \mu$.
Thm 14 (CLT): $X_1, \ldots, X_n$ iid, $\mathbb{E}(X_i) = \mu$ $\mathrm{Var}(X_i) = \sigma^2$
$\implies \sqrt{n}(\overline{X}_n - \mu)/\sigma \rightsquigarrow \mathcal{N}(0,1)$
$\implies \overline{X}_n \rightsquigarrow \mathcal{N}(\mu, \sigma^2/n)$
$\implies \sqrt{n}(\overline{X}_n - \mu)/S_n \rightsquigarrow \mathcal{N}(0,1)$
Thm 18 (delta method): If $\sqrt{n}(Y_n - \mu)/\sigma \rightsquigarrow \mathcal{N}(0,1)$, $g'(\mu) \neq 0$
$\implies \sqrt{n}(g(Y_n) - g(\mu))/|g'(\mu)|\sigma \rightsquigarrow \mathcal{N}(0,1)$
ie $Y_n \approx \mathcal{N}(\mu, \sigma^2/n) \implies g(Y_n) \approx \mathcal{N}(g(\mu), g'(\mu)^2\sigma^2/n)$
Thm 18b (2nd order delta method):?? Should I include this?

Sufficiency

If $X_1, \ldots, X_n \sim p(x;\theta)$, T **sufficient** for θ if $p(x^n|t;\theta) = p(x^n|t)$.
Thm 9 (factorization): for $X^n \sim p(x;\theta)$, $T(X^n)$ sufficient for θ if the joint probability can be factorized as.

$$p(x^n;\theta) = h(x^n) \times g(t;\theta) \tag{18}$$

T is a **minimal sufficient statistic (MSS)** if T is sufficient and $T = g(U)$ for all other sufficient stats U.
Thm 15: T is a MSS if:

$$\frac{p(y^n;\theta)}{p(x^n;\theta)} \text{ constant in } \theta \iff T(y^n) = T(x^n) \tag{19}$$

Parametric Point Estimation

make sure i've defined: $\mathbb{E}_\theta(\hat{\theta})$, bias, sampling distro, standard error, $\hat{\theta}_n$ consistent.
Method of Moments: Define equations
(a) $(\sum_i X_i)/n = \mathbb{E}_{\hat{\theta}}(X_i)$
(b) $(\sum_i X_i^2)/n = \mathbb{E}_{\hat{\theta}}(X_i^2)$
(c) ...
And solve for $\hat{\theta}$.
Maximum Likelihood (MLE): The MLE is

$$\hat{\theta} = \arg\min_\theta L(\theta) = \arg\min_\theta l(\theta) \tag{20}$$

Often suffices to solve for θ in $\frac{\partial l(\theta)}{\partial \theta} = 0$. The MLE is **equivariant** $\implies$ if $\eta = g(\theta)$ then $\hat{\eta} = g(\theta)$.
Bayes Estimation: For prior $\pi(\theta)$, choose

$$\hat{\theta} = \mathbb{E}(\theta|x^n) = \int \theta\pi(\theta|x^n)d\pi \tag{21}$$

Mean Squared Error (MSE): The MSE is

$$\text{MSE} = \mathbb{E}(\hat{\theta}-\theta)^2 = \int (\hat{\theta}-\theta)^2 p(x^n;\theta)dx^n = \text{bias}(\hat{\theta})^2 + Var(\hat{\theta}) \tag{22}$$

Notes: **bias**$(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$. We say $\hat{\theta}$ is **consistent** if $\hat{\theta} = \hat{\theta}_n \xrightarrow{p} \theta$. The **standard error** of $\hat{\theta}$, se$(\hat{\theta})$, is the standard deviation of $\hat{\theta}$.
Ex (in class): MSE for normal.

Risks and Estimators

$L(\theta, \hat{\theta})$ is the **loss** of an estimator $\hat{\theta} = \hat{\theta}(x^n)$ for $x^n \sim p(x^n;\theta)$.
The **risk** of $\hat{\theta}$ is

$$R(\theta,\hat{\theta}) = \mathbb{E}[L(\theta,\hat{\theta})] = \int L(\theta,\hat{\theta})p(x^n;\theta)dx^n \tag{23}$$

When $L(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$, the risk is the MSE.
The **max risk** of $\hat{\theta}$ over a set $\theta \in \Theta$ is

$$\overline{R}(\hat{\theta}) = \sup_{\theta\in\Theta} R(\theta,\hat{\theta}) \tag{24}$$

The **minimax estimator** is

$$\hat{\theta} = \arg\inf_{\hat{\theta}} \overline{R}(\hat{\theta}) \tag{25}$$

The **Bayes risk** of $\hat{\theta}$ given a prior $\pi(\theta)$ is

$$B_\pi(\hat{\theta}) = \int R(\theta,\hat{\theta})\pi(\theta)d\theta \tag{26}$$

The **posterior risk** of $\hat{\theta}$ given a prior $\pi(\theta)$ is

$$r(\hat{\theta}|x^n) = \int L(\theta,\hat{\theta})\pi(\theta|x_1,\ldots,x_n)d\theta \tag{27}$$

where $\pi(\theta|x^n) = \frac{\mathbb{P}(x^n;\theta)\pi(\theta)}{m(x^n)}$ is the posterior over θ.
The **Bayes estimator** is

$$\hat{\theta} = \arg\inf_{\hat{\theta}} B_\pi(\hat{\theta}) = \arg\inf_{\hat{\theta}} r(\hat{\theta}|x^n) \tag{28}$$

which equals $\mathbb{E}(\theta|x^n)$ when $L(\theta,\hat{\theta}) = (\theta - \hat{\theta})^2$.
Thm 10: If $\hat{\theta}$ is a Bayes estimator for some prior π and $R(\theta,\hat{\theta})$ is constant, then $\hat{\theta}$ is a minimax estimator.
Note: The MLE is approximately minimax (as n increases, if dimension of the parameter is fixed).