Probability Review

Concepts you should already know:

- Probability space
- Independent events
- Random variables
- Independent RVs
- Distribution, probability mass function, density
- Expectation (definition, mean, covariance, etc.)
Some basic R.V.s that will appear

Bernoulli \( p \): \( X \in \{0, 1\} \)

\[ P(X = 1) = p, \quad P(X = 0) = 1 - p \]

Multinomial \((k, (p_1, p_2, \ldots, p_n))\)

where \( p_j > 0 \), \( \sum_{j=1}^{n} p_j = 1 \)

\( X \in \mathbb{N}^n \), \( \sum_{j=1}^{n} X_j = k \)

\[ P(X = (m_1, \ldots, m_n)) = \frac{k!}{m_1! m_2! \cdots m_n!} \times p_1^{m_1} p_2^{m_2} \cdots p_n^{m_n} \]

Uniform \((a, b)\): \( a < b \)

\( X \in \mathbb{R} \)

density \( f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in (a, b) \\ 0, & \text{otherwise} \end{cases} \)

cdf \( F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases} \)
Multivariate Gaussian \((N(m, C))\):

\[
m \in \mathbb{R}^n \quad C \in \mathbb{R}^{n \times n} \quad \text{SPD}
\]

\[
f_X(x) = \frac{e^{-(x-m)^T C^{-1} (x-m)/2}}{(2\pi)^{n/2} \sqrt{\det C}}
\]

\[
\mathbb{E}[X] = m \quad \text{Cov}(X) = C
\]
You should also have seen a basic definition of conditional expectation

e.g.

\[ E[h(X,Y) \mid X=x] = \sum_y h(x,y) \frac{p(x,y)}{\sum_y p(x,y)} \]

when \( X \) and \( Y \) are discrete R.V.'s

\( \Psi \) joint prob mass fn \( p(x,y) \)

and \( E[h(X,Y) \mid X] = G(X) \) where

\( G(x) = E[h(X,Y) \mid X=x] \)

(replace \( p \) by the joint dusty and \( \sum_y \) by \( \text{Sdag} \) if \( X \) and \( Y \) are cont)

I will use a more general definition

\( E[X \mid \mathcal{F}] \) is the unique (up to events of prob 0) R.V. such that

\[ E[E[X \mid \mathcal{F}] \mid A] = E[X \mid A] \]

for all \( A \in \mathcal{F} \)
here $\mathcal{G}$ is a $\sigma$-algebra.

If you don't know what that is, just think of it as information.

$Y$ measurable w.r.t. $\mathcal{G}$ means that $\mathcal{G}$ contains all information about the value of $Y$, i.e. given the information in $\mathcal{G}$, $Y$ is constant.

So if $Y$ is measurable w.r.t. $\mathcal{G}$ then

$$E[X|Y] = E[E[X|Y]|Y]$$

(in particular $E[Y|\mathcal{G}] = Y$)

In the case that $\mathcal{G}$ is the smallest $\sigma$-algebra with respect to which $Y$ is measurable,

then $E[X|Y]$ coincides with

the definition of $E[X|Y]$ above.
Tower property of conditional expectation:

if $G \subset F$ (i.e. if $F$ contains all information in $G$)

then $E[E[X] \mid G] = E[E[X] \mid F] = E[X]$ |

for example $E[E[X] \mid Y] = E[E[X] \mid Y] = E[X]$

Basic inequality:

Markov: $I \Rightarrow P(X > 0) = 1$ and $E[X] < \infty$ then

$P(X > t) \leq \frac{E[X]}{t}$ $\forall t > 0$

Chebyshev: $I \Rightarrow \mu = E[X]$ and $\sigma^2 = \text{Var}(X)$

then $P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}$
Chernoff: \( \forall t \geq 0 \) and any \( a \in \mathbb{R} \)
\[
P(X \geq a) \leq \frac{E[etX]}{e^{ta}}
\]
proof by Markov
\[
P(X \geq a) = P(etX \geq e^{ta})
\]
\[
\leq \frac{E[etX]}{e^{ta}}
\]
Now suppose \( Y_1, \ldots, Y_n \) are i.i.d.
and \( X = \frac{1}{n} \sum Y_j \)
by Chebyshev
\[
P \left( |X - \mu| \geq t \right) \leq \frac{\text{Var}(X)}{t^2}
\]
\[
= \frac{\text{Var}(Y_1)}{n \cdot t^2}
\]
by Chernoff.

\[ P(X - \mu \geq a) = \mathbb{E} \left[ e^{\frac{n}{\mu} \sum_{j=1}^{n} Y_j} - 1 \right] \]

\[ = e^{-n ta} \mathbb{E} \left[ e^{t Y_j} \right]^n \]

\[ = e^{-n \left( t(\mu+a) - \log \mathbb{E} \left[ e^{t Y_j} \right] \right)} \]

\[ = e^{-n \chi(a)} \]

where \( \chi(a) = \inf_t \left( t(\mu+a) - \log \mathbb{E} \left[ e^{t Y_j} \right] \right) \)

Convergence of sequences of R.V.'s

\[ X_n \to X \text{ a.s. if } \lim_{n \to \infty} P(\lim_{n \to \infty} X_n = X) = 1 \]

\[ X_n \to X \text{ if } \forall \varepsilon > 0 \]

\[ \lim_{n \to \infty} P(\mid X_n - X \mid > \varepsilon) = 0 \]
\[
X_n \xrightarrow{d} X \quad \text{if} \quad F_{X_n}(x) \rightarrow F_X(x)
\]

where \( F \) is cant at \( x \)

or \[
E[g(X_n)] \rightarrow E[g(X)]
\]

\( \forall \) cant, bound \( g \)

\[
X_n \xrightarrow{L^2} X \quad \text{if} \quad E[|X_n - X|^2] \rightarrow 0
\]

\text{Summary:} \quad X_n \rightarrow X \ a.s. \Rightarrow X_n \xrightarrow{P} X

\Rightarrow X_n \xrightarrow{d} X

\[
X_n \xrightarrow{d} c \Rightarrow X_n \xrightarrow{P} c
\]

\[
X_n \xrightarrow{L^2} X \Rightarrow X_n \xrightarrow{P} X
\]
the Delta Method

\[ \sqrt{n}(Y_n - \theta) \overset{d}{\rightarrow} N(0, \sigma^2) \quad \text{and} \quad g'(\theta) \neq 0 \quad \text{then} \]

\[ \sqrt{n} \left( g(Y_n) - g(\theta) \right) \overset{d}{\rightarrow} N(0, \sigma^2 g'(\theta)^2) \]

---

**Basic Monte Carlo**

In MC we approximate \( E[f(X)] = \int f(x) \pi(dx) \) by a weighted sum

\[ \sum_{i=1}^{n} W(i) f(X(i)) \]

where \( W(i) \geq 0 \) and \( X(i) \) are random
For example, if we can draw
\[ X^{(1)}, X^{(2)}, \ldots, X^{(n)} \text{ i.i.d. from } \pi \]
then
\[
\text{Var} \left( \frac{1}{n} \sum_{j=1}^{n} f(X^{(j)}) \right) = \frac{\text{Var}(f(X^{(i)}))}{n} \rightarrow 0 \text{ as } n \rightarrow \infty
\]

In particular, since \( \|f\|_{\infty} \leq 1 \Rightarrow \text{Var} f \leq 1 \),

\[
\sup_{\|f\|_{\infty} \leq 1} E \left[ \left( \frac{1}{n} \sum_{j=1}^{n} f(X^{(j)}) - \int f(x) \right)^2 \right] \leq \frac{1}{n}
\]

No deterministic quadrature scheme can match

Note that, in principle we can approximate

any integral \( \int f(x)dx \) using MC

by writing \( \int f(x)dx = \int \left( \frac{f(x)}{\rho(x)} \right) \rho(x)dx \)
This is an example of importance sampling given target density $\pi$ and a reference density $\rho \gg \pi$ we can write

$$
\int f \pi(dx) = \int \left( \frac{f(x)\pi(x)}{\rho(x)} \right) \rho(dx)
$$

So if $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ i.i.d $\sim \rho$

then

$$
\int f \pi(dx) \approx \frac{1}{n} \sum_{j=1}^{n} \frac{\pi(X^{(j)})}{\rho(X^{(j)})}
$$

Useful when we can't sample directly from $\pi$

What is the variance of this estimate

$$
\text{Var} \left( \frac{1}{n} \sum_{j=1}^{n} \frac{\pi(X^{(j)})}{\rho(X^{(j)})} f(X^{(j)}) \right) = \frac{\text{Var} \left( \frac{\pi}{\rho} f(X^{(i)}) \right)}{n}
$$

$$
\text{Var} \left( \frac{\pi}{\rho} f(X^{(i)}) \right) = \int \left( \frac{\pi}{\rho} \right)^2 f^2(x) \rho(dx) - \left( \int \frac{\pi}{\rho} f(x) \rho(dx) \right)^2
$$
For \( f = 1 \) we get

\[
V_a \left( \frac{\pi}{\theta} \right) = \int (1 - \frac{\pi}{\theta})^2 \rho(dx) = \chi^2 (\pi \parallel \theta)
\]

So the variance of the weights will be large when \( \theta \) is very close to \( \pi \).

What is the best choice of \( \theta \)?

\[
V_a \left( \frac{\theta}{\pi} \phi (X^{(n)}) \right)
\]

\[
\geq \left( \int \theta \pi (dx) \right)^2 - \left( \int \phi (dx) \right)^2
\]

by Jensen's

This is achieved by \( \theta (x) = \frac{|f(x)| \pi (x)}{\int |f(x)| \pi (x) dx} \)

(In most cases we can't sample from \( \pi(x) \))

IS variance can be much lower than \( \theta = \pi \)

It can also be much higher if we use a poor choice of \( \theta \) ... particularly true i high d
Note that in IS we need to be able to evaluate \( \frac{f}{g} \). This may be impossible because usually we only know some

\[ p \propto n \]

Suppose \( r p \) and \( p \propto n \)

Instead we can use the estimate

\[
\frac{1}{n} \sum_{j=1}^{n} \frac{p(X^{(i)})}{g(X^{(i)})} \quad \rightarrow \quad \int \frac{p}{g} \, dx = \left( \frac{f}{g} \right)_{\text{est}}
\]

\[
\frac{1}{n} \sum_{j=1}^{n} \frac{g(X^{(i)})}{r(X^{(i)})} \quad \rightarrow \quad \int \frac{p}{r} \, dx = \frac{Sp}{Sr}
\]

This estimate has a bias

\[
E \left[ \frac{1}{n} \sum_{j=1}^{n} \frac{p(X^{(i)})}{g(X^{(i)})} \right] \neq E \left[ \frac{1}{n} \sum_{j=1}^{n} \frac{f}{g} \right]
\]

\[
= \int p \, d\lambda(x)
\]

but

\[
\frac{x}{y} = \frac{\bar{x}}{\bar{y}} + \frac{1}{y} (x-\bar{x}) - \frac{\bar{x}}{y^2} (y-\bar{y})
\]

\[
+ \frac{1}{y^2} \frac{(y-\bar{y})(x-\bar{x})}{2} + \frac{3\bar{x}}{2y^3} (y-\bar{y})^2 + O(n^{-\frac{3}{2}})
\]
So bias is $O\left(\frac{1}{n}\right)$

Smaller than statistical error when $n$ is large

We usually cannot sample directly from $\pi$.

But we can almost always construct a correlated sequence $X^{(1)}, X^{(n)}, \ldots, X^{(n)}$ with $X^{(e)} \to \pi$ and

$$\frac{1}{n} \sum_{t=1}^{n} f(X^{(t)}) \to \int f \pi \, dx$$

$X^{(t)}$ is a Markov Process, i.e.

if $\mathcal{F}_t$ is the $\sigma$-algebra generated by $X^{(1)}, X^{(2)}, \ldots, X^{(t)}$

Then for $k \geq t$

$$P(X^{(k)} \in B \mid \mathcal{F}_t) = P(X^{(k)} \in B \mid X^{(t)})$$

i.e. If I know the state today, I don't need to know the past
e.g. \[ X^{(k+1)} = (I - A)X^{(k)} + b \]

is a (deterministic) Markov process

but if the recursion depended on \( X^{(k-1)} \) too then it would not be (though we might be able to rewrite it as a Markov process in a higher dimensional space)

Note that if the \( X^{(i)} \) are correlated let's assume \( X^{(i)} \sim \pi \) and \( \bar{F} = \int f(x) \, dx \)

\[
\text{Var} \left( \frac{1}{n} \sum_{t=1}^{n} f(X^{(t)}) \right) = \frac{\text{Var}(f(X^{(i)}))}{n} + \frac{1}{n^2} \sum_{t,j,k=1}^{n} \text{Cov}(f(X^{(t)}), f(X^{(k)}))
\]

\[
= \frac{\text{Var}(f(X^{(i)}))}{n} \left( 1 + 2 \sum_{t=1}^{n} \sum_{k=1}^{n-t} \frac{\text{Cov}(f(X^{(k)}), f(X^{(k+t)}))}{n} \right)
\]

\[
= \frac{\text{Var}(f(X^{(i)}))}{n} \left( 1 + 2 \sum_{k=1}^{n} \frac{n-k}{n} \text{Cov}(f(X^{(k)}), f(X^{(k+t)})) \right)
\]
So we expect

\[ n \text{Var} \left( \frac{1}{n} \sum_{t=1}^{n} f(X^{(t)}) \right) \rightarrow \text{Var}(f(X^{(0)})) \text{ as } n \rightarrow \infty \]

\[ \nu \bigg/ \tau_f = 1 + 2 \sum_{t=1}^{\infty} \text{corr}(f(X^{(t)}), f(X^{(t+t)})) \]

In fact, suppose \( X^{(k)} \) is uniformly ergodic,

e.g. if \( \sup_{x} \| P^{+}(x, \cdot) - \pi(\cdot) \|_{TV} \leq C 2^{-k} \text{ o}(\text{a.s.}) \)

\[ \mathbb{P}(X, A) = \mathbb{P}_{x}(X^{(k)} \in A) \]

\[ \| \eta - \nu \|_{TV} = \int \frac{d\eta}{d\nu(\eta + \nu)} - \frac{d\nu}{d\nu(\eta + \nu)} d(\eta + \nu) \]

and if \( 0 < \tau_f < \infty \)

\[ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} f(X^{(t)}) - \bar{f} \rightarrow \mathcal{N}(0, \text{Var}(f(X^{(0)}))) \]

We can estimate \( \tau_f \) in practice.

To report approximate error bars.
Finite state case:

\[ S \in \mathbb{R}^{n \times n} \quad S_{ij} \geq 0 \quad \sum_j S_{ij} = 1 \]

\[ P \left( X^{(t+1)} = j \mid X^{(t)} = i \right) = S_{ij} \]

\( \pi \) is the invariant probability vector if

\[ \forall j \quad \sum_i \pi_i \cdot P \left( X^{(t+1)} = j \mid X^{(t)} = i \right) = \pi_j \]

i.e. \( \pi^T S = \pi^T \)

Note that \( S \) is always a \( l_1 \) contraction

Let \( (V^+)_i = \max \{ v_i, 0 \} \) and \( (V^-)_i = \max \{ -v_i, 0 \} \)

so \( V = V^+ - V^- \) and \( \| Vi \| = (V^+)_i + (V^-)_i \)

Then \( \| V^T S \|_1 \leq \| V^T S \|_1 + \| V^- S \|_1 \)

\[ = \| V^+ S 1 \|_1 + \| V^- S 1 \|_1 \]

\[ = \| V^+ 1 \|_1 + \| V^- 1 \|_1 \]

\[ = \| V \|_1 \]

but not strict
If $S$ is irreducible and aperiodic then
\[
\max_i \| e_i S^t - \pi \|_1 \leq C 2^t \quad \text{for } \pi \in \{0,1\}
\]

$S$ is irreducible if $\forall i,j \exists t \text{ s.t. } (S^t)_{ij} > 0$

The chain is called aperiodic if for all $i$

the greatest common divisor of

\[
\{ t > 1 : (S^t)_{ii} > 0 \} \text{ is } 1
\]

Irreducible and aperiodic implies $\exists k \text{ s.t.}$

\[
(S^k)_{ij} > 0 \quad \forall i,j
\]

In this case, if $\eta \in \mathbb{R}^n \setminus \{0\}$

then \[
(\eta^T S^k)_{j} = \sum_i \eta_i (S^k)_{ij} > (\min_{i,j} S^k_{ij}) \| \eta \|_1
\]

i.e. \[
\eta^T S^k > \delta \| \eta \|_1 \frac{1}{n}
\]
Suppose $V^T 1 = 0$, so $\mathcal{M} = V^T 1 = \|v_U 1\|_1 - \|v_L 1\|_1$.

And $\|u_1\|_1 = \|v_U 1\|_1 + \|u_L 1\|_1$, so $\|v_U 1\|_1 = \|u_1\|_1 = \frac{\|u_1\|_1}{2}$.

Thus $\|v^T S^k 1\|_1 \leq \|V^T S^k 1 - \frac{S}{2}\|_1 + \|V^T S^k 1 - \frac{S}{2}\|_1 \geq 0$.

So $V^T S^k 1 + V^T S^k 1 - \frac{S}{2} \|u_1\|_1$,

$= V^T 1 + V^T 1 - \frac{S}{2} \|u_1\|_1$,

$= (1 - \frac{1}{2}) \|u_1\|_1$.

So $\|V^T S^k 1 - \pi^T 1\|_1 = \|V^T 1 - \pi^T 1\|_1$,

$\leq (1 - \frac{1}{8}) \|V^T 1 - \pi^T 1\|_1$,

$\leq (1 - \frac{1}{8}) \|V^T 1 - \pi^T 1\|_1$.

Choose $\lambda = (1 - \frac{1}{8}) \frac{1}{1 - \frac{1}{8}}$, $C = \frac{1}{1 - \frac{1}{8}}$.

$\|V^T S^k 1 - \pi^T 1\|_1 \leq C \lambda^k$. 