

Lecture 4: Variance, Covariance, and Common Discrete Distributions

EECS 126 (UC Berkeley)

Spring 2026 — January 29, 2026

Reading: Bertsekas & Tsitsiklis, Ch. 2; §4.2–4.3. Ch. 3 next week.

These are my personal notes from Professor Kannan Ramchandran's EECS 126 (Probability and Random Processes) course at UC Berkeley.

Last lecture we defined expectation and variance for a single random variable. In this lecture we study how variance behaves for *sums* of random variables, which leads to covariance and the correlation coefficient. We then compute the mean and variance of the four workhorse discrete distributions (Bernoulli, binomial, geometric, Poisson), apply these to the coupon collector problem, prove the memoryless property of the geometric distribution, study how Poisson random variables merge and split, and finally introduce conditional expectation.

4.1 Variance of Sums and Covariance

Recall from Lecture 3 the two forms of the variance,

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2, \quad (4.1)$$

and the scaling property

$$\text{Var}(aX + c) = a^2 \text{Var}(X). \quad (4.2)$$

Shifts by a constant c are not recorded in variance calculations: shifting every outcome by c moves the mean by c as well, so the deviations $X - \mathbb{E}[X]$ are unchanged.

The natural next question is: what is the variance of a *sum* $X + Y$? Unlike the mean, the answer depends on how X and Y interact.

Theorem 4.1 (Variance of a sum)

For any random variables X, Y with finite variances,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]). \quad (4.3)$$

Proof.

$$\begin{aligned} \text{Var}(X + Y) &= \mathbb{E}[(X + Y)^2] - (\mathbb{E}[X + Y])^2 \\ &= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - (\mathbb{E}[X])^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - (\mathbb{E}[Y])^2 \\ &= \underbrace{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}_{\text{Var}(X)} + \underbrace{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}_{\text{Var}(Y)} + 2\mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] \\ &= \text{Var}(X) + \text{Var}(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]). \quad \square \end{aligned}$$

No independence was used. The derivation above holds for *any* X and Y . Independence only enters at the very end, where it makes the cross term vanish (Corollary 4.1 below).

The cross term in (4.3) is important enough to get its own name.

Definition 4.1 (Covariance)

The **covariance** of random variables X and Y is

$$\text{Cov}(X, Y) \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \quad (4.4)$$

In particular, $\text{Cov}(X, X) = \text{Var}(X)$. With this notation,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2 \text{Cov}(X, Y). \quad (4.5)$$

Equivalence of the two forms. Write $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$, which are constants. Expanding the product and using linearity,

$$\begin{aligned} \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] &= \mathbb{E}[XY - \mu_Y X - \mu_X Y + \mu_X \mu_Y] \\ &= \mathbb{E}[XY] - \mu_Y \mathbb{E}[X] - \mu_X \mathbb{E}[Y] + \mu_X \mu_Y \\ &= \mathbb{E}[XY] - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y \\ &= \mathbb{E}[XY] - \mu_X \mu_Y. \end{aligned} \quad \square$$

Interpreting the sign. $\text{Cov}(X, Y) > 0$ means X and Y tend to be above their means together (and below their means together). $\text{Cov}(X, Y) < 0$ means when one is above its mean, the other tends to be below. For a sum, positive covariance *inflates* the spread and negative covariance *dampens* it.

Corollary 4.1 (Independent summands)

If $X \perp Y$, then $\text{Cov}(X, Y) = 0$, and therefore

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y). \quad (4.6)$$

Proof. In Lecture 3 we proved that $X \perp Y$ implies $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Hence $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$, and the cross term in (4.3) vanishes. $\square$

Corollary 4.2 (Variance of a sum of many random variables)

For random variables $X_1, \dots, X_n$ with finite variances,

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j). \quad (4.7)$$

In particular, if the X_i are independent (in fact, *pairwise uncorrelated* is enough), then

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i). \quad (4.8)$$

Proof. Let $\mu_i = \mathbb{E}[X_i]$. Since $\mathbb{E}[\sum_i X_i] = \sum_i \mu_i$, the centered sum is $\sum_i (X_i - \mu_i)$. Squaring a sum gives a double sum over all pairs (i, j) :

$$\begin{aligned} \text{Var}\left(\sum_i X_i\right) &= \mathbb{E}\left[\left(\sum_i (X_i - \mu_i)\right)^2\right] = \mathbb{E}\left[\sum_i \sum_j (X_i - \mu_i)(X_j - \mu_j)\right] \\ &= \sum_i \sum_j \text{Cov}(X_i, X_j). \end{aligned} \quad (4.9)$$

The diagonal terms $i = j$ give $\text{Cov}(X_i, X_i) = \text{Var}(X_i)$. The off-diagonal terms come in symmetric pairs, $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$, so they contribute $2 \sum_{i < j} \text{Cov}(X_i, X_j)$. If every pair is uncorrelated, all off-diagonal terms are zero. $\square$

Uncorrelated does not imply independent. Covariance only detects *linear* dependence. For example, let X be uniform on $\{-1, 0, 1\}$ and $Y = X^2$. Then $\mathbb{E}[X] = 0$ and $\mathbb{E}[XY] = \mathbb{E}[X^3] = \frac{1}{3}(-1) + \frac{1}{3}(0) + \frac{1}{3}(1) = 0$, so $\text{Cov}(X, Y) = 0 - 0 \cdot \mathbb{E}[Y] = 0$. But Y is completely determined by X : for instance $\mathbb{P}(Y = 0 \mid X = 0) = 1 \neq \frac{1}{3} = \mathbb{P}(Y = 0)$. This was a question Ramchandran asked us at end of previous lecture.

4.2 The Correlation Coefficient

Covariance has units (units of X times units of Y), so its size alone does not tell us how *strongly* two variables are related. Normalizing by the standard deviations removes the units.

Definition 4.2 (Correlation coefficient)

For random variables X, Y with $\text{Var}(X) > 0$ and $\text{Var}(Y) > 0$, the **correlation coefficient** is

$$\rho(X, Y) \triangleq \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}}. \quad (4.10)$$

Ramchandran’s vector analogy. Think of random variables as vectors, with “dot product” $\langle U, V \rangle = \mathbb{E}[UV]$ and “length” $\sqrt{\mathbb{E}[U^2]}$. For the centered variables $U = X - \mathbb{E}[X]$ and $V = Y - \mathbb{E}[Y]$:

- the dot product is $\mathbb{E}[UV] = \text{Cov}(X, Y)$;
- the lengths are $\sqrt{\mathbb{E}[U^2]} = \sqrt{\text{Var}(X)}$ and $\sqrt{\mathbb{E}[V^2]} = \sqrt{\text{Var}(Y)}$.

So ρ plays the role of $\cos \theta = \frac{u \cdot v}{|u| |v|}$, the cosine of the angle between two vectors. For ordinary vectors, the Cauchy–Schwarz inequality $|u \cdot v| \leq |u| |v|$ forces $-1 \leq \cos \theta \leq 1$. The same inequality, proved for random variables below, forces $-1 \leq \rho \leq 1$.

Lemma 4.1 (Cauchy–Schwarz inequality for random variables)

For any random variables U, V with $\mathbb{E}[U^2] < \infty$ and $\mathbb{E}[V^2] < \infty$,

$$(\mathbb{E}[UV])^2 \leq \mathbb{E}[U^2] \mathbb{E}[V^2]. \quad (4.11)$$

Proof. **Case 1:** $\mathbb{E}[V^2] = 0$. Since $V^2 \geq 0$ and its mean is 0, we must have $V = 0$ with probability 1. Then $UV = 0$ with probability 1, so $\mathbb{E}[UV] = 0$ and both sides equal 0.

Case 2: $\mathbb{E}[V^2] > 0$. For every real number t , the random variable $(U - tV)^2$ is nonnegative, so its expectation is nonnegative. Expanding with linearity:

$$0 \leq \mathbb{E}[(U - tV)^2] = \mathbb{E}[U^2] - 2t\mathbb{E}[UV] + t^2\mathbb{E}[V^2]. \quad (4.12)$$

This holds for every t , so choose the t that minimizes the right-hand side, namely $t^* = \mathbb{E}[UV]/\mathbb{E}[V^2]$. Substituting,

$$\begin{aligned} 0 &\leq \mathbb{E}[U^2] - 2 \frac{\mathbb{E}[UV]}{\mathbb{E}[V^2]} \mathbb{E}[UV] + \frac{(\mathbb{E}[UV])^2}{(\mathbb{E}[V^2])^2} \mathbb{E}[V^2] \\ &= \mathbb{E}[U^2] - 2 \frac{(\mathbb{E}[UV])^2}{\mathbb{E}[V^2]} + \frac{(\mathbb{E}[UV])^2}{\mathbb{E}[V^2]} \\ &= \mathbb{E}[U^2] - \frac{(\mathbb{E}[UV])^2}{\mathbb{E}[V^2]}. \end{aligned} \quad (4.13)$$

Multiplying both sides by $\mathbb{E}[V^2] > 0$ gives $(\mathbb{E}[UV])^2 \leq \mathbb{E}[U^2] \mathbb{E}[V^2]$. $\square$

Theorem 4.2 (Range of the correlation coefficient)

Whenever $\rho(X, Y)$ is defined,

$$-1 \leq \rho(X, Y) \leq 1. \quad (4.14)$$

Proof. Apply Cauchy–Schwarz to the centered variables $U = X - \mathbb{E}[X]$ and $V = Y - \mathbb{E}[Y]$. By definition, $\mathbb{E}[UV] = \text{Cov}(X, Y)$, $\mathbb{E}[U^2] = \text{Var}(X)$, and $\mathbb{E}[V^2] = \text{Var}(Y)$, so

$$(\text{Cov}(X, Y))^2 \leq \text{Var}(X) \text{Var}(Y). \quad (4.15)$$

Taking square roots, $|\text{Cov}(X, Y)| \leq \sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}$. Dividing by the (positive) right-hand side gives $|\rho(X, Y)| \leq 1$. $\square$

Extreme and zero values. $|\rho| = 1$ exactly when equality holds in Cauchy–Schwarz, which happens when $U - t^*V = 0$ with probability 1, i.e. when Y is an exact linear function $Y = aX + b$ of X (with $\rho = +1$ if $a > 0$ and $\rho = -1$ if $a < 0$). At the other end, $\rho = 0$ means *uncorrelated*, which, as we just saw, is weaker than independent.

4.3 Mean and Variance of Common Distributions

We now compute $\mathbb{E}[X]$ and $\text{Var}(X)$ for the four distributions introduced in Lecture 3. A recurring theme: the direct sums are often messy, and a better *representation* of the random variable makes the computation easy.

Bernoulli

Theorem 4.3 (Bernoulli mean and variance)

If $X \sim \text{Bernoulli}(p)$, then

$$\mathbb{E}[X] = p, \quad \text{Var}(X) = p(1 - p). \quad (4.16)$$

Proof. X takes only the values 0 and 1, with $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$.

$$\mathbb{E}[X] = 1 \cdot \mathbb{P}(X = 1) + 0 \cdot \mathbb{P}(X = 0) = 1 \cdot p + 0 = p. \quad (4.17)$$

For the second moment, by LOTUS,

$$\mathbb{E}[X^2] = \sum_x x^2 \mathbb{P}(X = x) = 1^2 \cdot \mathbb{P}(X = 1) + 0^2 \cdot \mathbb{P}(X = 0) = p. \quad (4.18)$$

Therefore

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1 - p). \quad (4.19)$$

□

Binomial

Theorem 4.4 (Binomial mean and variance)

If $X \sim \text{Binomial}(n, p)$, then

$$\mathbb{E}[X] = np, \quad \text{Var}(X) = np(1 - p). \quad (4.20)$$

The direct approach is

$$\mathbb{E}[X] = \sum_{x=0}^n x \cdot \mathbb{P}(X = x) = \sum_{x=0}^n x \binom{n}{x} p^x (1 - p)^{n-x}, \quad (4.21)$$

which is doable but unpleasant. Instead, think of the binomial as a *sum of n Bernoulli trials*.

Proof. Let $X_i = \mathbf{1}\{\text{trial } i \text{ is a success}\}$ for $i = 1, \dots, n$. Each $X_i \sim \text{Bernoulli}(p)$, the X_i are independent (the trials are independent), and the total number of successes is

$$X = X_1 + X_2 + \dots + X_n. \quad (4.22)$$

Mean. By linearity of expectation (which needs no independence),

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n p = np. \quad (4.23)$$

Variance. Because the X_i are independent, Corollary 4.1 lets us add the variances:

$$\text{Var}(X) = \sum_{i=1}^n \text{Var}(X_i) = \sum_{i=1}^n p(1 - p) = np(1 - p). \quad (4.24)$$

□

Applying direct sum gives us the same result: Using the identity $x \binom{n}{x} = n \binom{n-1}{x-1}$ for

$x \geq 1$ (the $x = 0$ term is zero),

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=1}^n n \binom{n-1}{x-1} p^x (1-p)^{n-x} = np \sum_{x=1}^n \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)} \\ &= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j} = np (p + (1-p))^{n-1} = np,\end{aligned}\quad (4.25)$$

where we substituted $j = x - 1$ and used the binomial theorem. The indicator method reaches the same answer with far less algebra.

Geometric

Recall that $X \sim \text{Geometric}(p)$ counts the number of independent trials up to and including the first success:

$$p_X(k) = \mathbb{P}(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, 3, \dots \quad (4.26)$$

Reading the PMF. Here k marks the trial on which the first Success occurs ($k = S$). So $p_X(k)$ is the probability that the first success happens on trial k : the first $k - 1$ trials are failures (probability $(1-p)^{k-1}$) and trial k is a success (probability p). Note that $p_X(k)$ is *not* the probability of k successes; a geometric random variable always records exactly one success, and k counts how long we waited for it.

Definition 4.3 (Survival function)

For an integer-valued random variable X , the **survival function** (or *tail*) is $k \mapsto \mathbb{P}(X > k)$. For $X \sim \text{Geometric}(p)$,

$$\mathbb{P}(X > k) = (1-p)^k, \quad k = 0, 1, 2, \dots \quad (4.27)$$

Proof. The event $\{X > k\}$ means “no success in the first k trials,” i.e. we tried k times and failed every time. The trials are independent and each fails with probability $1 - p$, so $\mathbb{P}(X > k) = \underbrace{(1-p) \cdots (1-p)}_{k \text{ times}} = (1-p)^k$. $\square$

Theorem 4.5 (Geometric mean)

If $X \sim \text{Geometric}(p)$ with $0 < p \leq 1$, then

$$\mathbb{E}[X] = \frac{1}{p}. \quad (4.28)$$

The direct sum $\mathbb{E}[X] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p$ requires a trick. A better method is the tail sum formula from Lecture 3, which applies because X takes values in the nonnegative integers.

Proof. By the tail sum formula,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k). \quad (4.29)$$

For integer-valued X , $\{X \geq k\} = \{X > k - 1\}$. Substituting $j = k - 1$ (so j runs from 0 to ∞),

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X > k - 1) = \sum_{j=0}^{\infty} \mathbb{P}(X > j). \quad (4.30)$$

Now plug in the survival function (4.27):

$$\mathbb{E}[X] = \sum_{j=0}^{\infty} (1 - p)^j. \quad (4.31)$$

This is a geometric series $\sum_{j \geq 0} x^j$ with ratio $x = 1 - p$. Since $0 < p \leq 1$ we have $|x| < 1$, so the series converges to $1/(1 - x)$:

$$\mathbb{E}[X] = \frac{1}{1 - (1 - p)} = \frac{1}{p}. \quad (4.32)$$

□

Exam note: tail sum formula will be on exam. For a random variable X taking values in $\{0, 1, 2, \dots\}$,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) = \sum_{k=0}^{\infty} \mathbb{P}(X > k). \quad (4.33)$$

It expresses the mean as the sum of the tail probabilities. It follows from writing $X = \sum_{k=1}^{\infty} \mathbf{1}\{X \geq k\}$ and taking expectations of both sides.

The direct sum, for comparison. Differentiating the geometric series $\sum_{k \geq 0} x^k = \frac{1}{1-x}$ term by term (valid for $|x| < 1$) gives $\sum_{k \geq 1} kx^{k-1} = \frac{1}{(1-x)^2}$. With $x = 1 - p$,

$$\mathbb{E}[X] = p \sum_{k=1}^{\infty} k(1 - p)^{k-1} = p \cdot \frac{1}{(1 - (1 - p))^2} = p \cdot \frac{1}{p^2} = \frac{1}{p}. \quad (4.34)$$

The variance of the geometric is derived in Section 4.5, using the memoryless property.

Poisson

Theorem 4.6 (Poisson mean and variance)

If $X \sim \text{Poisson}(\lambda)$, so that $p_X(k) = \frac{e^{-\lambda} \lambda^k}{k!}$ for $k = 0, 1, 2, \dots$, then

$$\mathbb{E}[X] = \lambda, \quad \text{Var}(X) = \lambda. \quad (4.35)$$

Both computations rely on the Taylor series of the exponential,

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots, \quad (4.36)$$

and on the cancellation $\frac{k}{k!} = \frac{1}{(k-1)!}$ for $k \geq 1$.

Proof of the mean.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=0}^{\infty} k \cdot \frac{e^{-\lambda} \lambda^k}{k!} \\
&= \sum_{k=1}^{\infty} k \cdot \frac{e^{-\lambda} \lambda^k}{k!} && \text{(the } k=0 \text{ term is 0)} \\
&= \sum_{k=1}^{\infty} \frac{e^{-\lambda} \lambda^k}{(k-1)!} && \text{(cancel } k \text{ against } k!) \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} && \text{(factor out } \lambda e^{-\lambda}) \\
&= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} && \text{(substitute } j = k-1) \\
&= \lambda e^{-\lambda} e^{\lambda} = \lambda. && \text{(Taylor series (4.36))} \quad \square
\end{aligned}$$

For the variance, computing $\mathbb{E}[X^2] = \sum k^2 \frac{e^{-\lambda} \lambda^k}{k!}$ directly is awkward because k^2 does not cancel cleanly against $k!$. The fix is the **factorial moment trick**: since $X^2 = X(X-1) + X$, linearity gives

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X], \quad (4.37)$$

and $k(k-1)$ cancels against $k!$ just as k did.

Proof of the variance. Step 1: the factorial moment.

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1) \frac{e^{-\lambda} \lambda^k}{k!} \\
&= \sum_{k=2}^{\infty} k(k-1) \frac{e^{-\lambda} \lambda^k}{k!} && \text{(the } k=0, 1 \text{ terms are 0)} \\
&= \sum_{k=2}^{\infty} \frac{e^{-\lambda} \lambda^k}{(k-2)!} && \text{(cancel } k(k-1) \text{ against } k!) \\
&= e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} && \text{(factor out } e^{-\lambda} \lambda^2) \\
&= e^{-\lambda} \lambda^2 e^{\lambda} = \lambda^2. && \text{(same trick: substitute } j = k-2) \quad (4.38)
\end{aligned}$$

Step 2: the second moment. By (4.37),

$$\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = \lambda^2 + \lambda. \quad (4.39)$$

Step 3: the variance.

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = (\lambda^2 + \lambda) - (\lambda)^2 = \lambda. \quad (4.40)$$

□

Mean equals variance (Poisson distribution). The Poisson is the only one of our four distributions whose mean and variance are equal. This is consistent with the Poisson being the limit of $\text{Binomial}(n, \lambda/n)$: the binomial variance is $n \cdot \frac{\lambda}{n} \left(1 - \frac{\lambda}{n}\right) = \lambda \left(1 - \frac{\lambda}{n}\right) \rightarrow \lambda$ as $n \rightarrow \infty$.

4.4 The Coupon Collector Problem

Example 4.1 (Coupon collector)

There are N different coupons you want to collect, found inside cereal boxes. Each box contains exactly one coupon, whose type is **uniformly random** among the N types and **independent** of all other boxes. You buy boxes until you have received all N distinct coupons. What is the expected number of boxes you must buy?

Let T be the total time (number of boxes) to collect all N coupons. Break the process into N stages. For $i = 1, \dots, N$, let

$$T_i = \# \text{ of boxes bought to obtain the } i\text{-th new coupon}, \quad (4.41)$$

counting from right after the $(i-1)$ -th new coupon was obtained. Then the total time is the sum of the stage times,

$$T = \sum_{i=1}^N T_i. \quad (4.42)$$

Each stage is geometric. During stage i we already hold $i-1$ distinct coupons, so $N - (i-1) = N - i + 1$ types would be new. Each box is independently a “success” (a new type) with probability

$$p_i = \frac{N - i + 1}{N}, \quad \text{so} \quad T_i \sim \text{Geometric}\left(\frac{N - i + 1}{N}\right). \quad (4.43)$$

- $i = 1$: $T_1 \sim \text{Geometric}\left(\frac{N-1+1}{N}\right) = \text{Geometric}(1)$, so $T_1 = 1$. The first box always gives a new coupon.
- $i = 2$: $T_2 \sim \text{Geometric}\left(\frac{N-2+1}{N}\right) = \text{Geometric}\left(\frac{N-1}{N}\right)$.
- $i = N$: $T_N \sim \text{Geometric}\left(\frac{1}{N}\right)$. The last coupon is the hardest to find.

Expected total. Since $\mathbb{E}[T_i] = 1/p_i = \frac{N}{N-i+1}$, linearity of expectation gives

$$\mathbb{E}[T] = \mathbb{E}\left[\sum_{i=1}^N T_i\right] = \sum_{i=1}^N \mathbb{E}[T_i] = \sum_{i=1}^N \frac{N}{N-i+1}. \quad (4.44)$$

Write out the terms, from $i = 1$ (denominator N) to $i = N$ (denominator 1):

$$\begin{aligned} \mathbb{E}[T] &= \frac{N}{N} + \frac{N}{N-1} + \frac{N}{N-2} + \cdots + \frac{N}{2} + \frac{N}{1} \\ &= N \left(\frac{1}{N} + \frac{1}{N-1} + \cdots + \frac{1}{2} + \frac{1}{1} \right) = N \sum_{j=1}^N \frac{1}{j}, \end{aligned} \quad (4.45)$$

where we substituted $j = N - i + 1$ (as i runs from 1 to N , j runs from N down to 1). The sum is the N -th **harmonic number** $H_N = \sum_{j=1}^N \frac{1}{j}$, so

$$\mathbb{E}[T] = N H_N. \quad (4.46)$$

Lemma 4.2 (Size of the harmonic numbers)

For every $N \geq 1$,

$$\ln(N+1) \leq H_N \leq 1 + \ln N, \quad (4.47)$$

and more precisely $H_N = \ln N + \gamma + o(1)$, where $\gamma \approx 0.5772$ is the **Euler–Mascheroni constant**.

Proof of the bounds. Compare each term $1/j$ with the area under the curve $1/x$. Since $1/x$ is decreasing, on the interval $[j, j+1]$ we have $1/x \leq 1/j$, and on $[j-1, j]$ we have $1/x \geq 1/j$. Hence

$$\int_j^{j+1} \frac{dx}{x} \leq \frac{1}{j} \leq \int_{j-1}^j \frac{dx}{x} \quad (j \geq 2 \text{ for the right inequality}). \quad (4.48)$$

Summing the left inequality over $j = 1, \dots, N$ gives $H_N \geq \int_1^{N+1} \frac{dx}{x} = \ln(N+1)$. Summing the right inequality over $j = 2, \dots, N$ and adding the $j = 1$ term gives $H_N \leq 1 + \int_1^N \frac{dx}{x} = 1 + \ln N$. (The sharper statement with γ is a classical result we take as given.) $\square$

Corollary 4.3 (Coupon collector asymptotics)

$$\mathbb{E}[T] = N H_N \approx N \ln N + \gamma N \approx N \ln N. \quad (4.49)$$

So we need to buy about a factor of $\ln N$ more boxes than the N coupons we are trying to collect.

Numbers and intuition. For $N = 50$, $H_{50} \approx 4.499$, so $\mathbb{E}[T] \approx 225$ boxes, about $4.5\times$ the number of coupon types. Most of the cost comes at the end: the final coupon alone takes $\mathbb{E}[T_N] = N$ boxes on average, while the entire first half of the collection takes fewer than N boxes in total ($\sum_{i \leq N/2} \frac{N}{N-i+1} \leq \frac{N}{2} \cdot 2 = N$). This $N \ln N$ behavior shows up whenever we need to “hit every bin” by random sampling: hashing, load balancing, and covering every node in a network.

4.5 The Memoryless Property

Theorem 4.7 (Memoryless property of the geometric)

Let $X \sim \text{Geometric}(p)$ with $0 < p < 1$. For all integers $s, t \geq 0$,

$$\mathbb{P}(X > s+t \mid X > s) = \mathbb{P}(X > t). \quad (4.50)$$

In words: given that we have already failed s times, the probability of failing t more times is the same as if we were starting over from scratch.

Proof. First note that $\mathbb{P}(X > s) = (1-p)^s > 0$, so the conditional probability is well defined. By the definition of conditional probability,

$$\mathbb{P}(X > s+t \mid X > s) = \frac{\mathbb{P}(\{X > s+t\} \cap \{X > s\})}{\mathbb{P}(X > s)}. \quad (4.51)$$

Since $s+t \geq s$, the event $\{X > s+t\}$ is contained in $\{X > s\}$ (if we have $s+t$ failures, we certainly have s failures), so the intersection is just $\{X > s+t\}$:

$$\mathbb{P}(X > s+t \mid X > s) = \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)}. \quad (4.52)$$

Both numerator and denominator are values of the geometric survival function (4.27):

$$\frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)} = \frac{(1-p)^{s+t}}{(1-p)^s} = (1-p)^t, \quad (4.53)$$

which is exactly the survival function of the geometric evaluated at t , i.e. $\mathbb{P}(X > t)$. $\square$

Two remarks on memorylessness.

1. **The remaining wait is again geometric.** Given $X > s$, the number of additional trials $X - s$ has the same distribution as X : for $k \geq 1$,

$$\mathbb{P}(X - s = k \mid X > s) = \frac{\mathbb{P}(X = s+k)}{\mathbb{P}(X > s)} = \frac{(1-p)^{s+k-1}p}{(1-p)^s} = (1-p)^{k-1}p. \quad (4.54)$$

This is the form we use in the derivations below.

2. **Uniqueness.** The geometric is the *only* memoryless distribution on $\{1, 2, 3, \dots\}$. Its continuous analog, the exponential distribution, will be the only memoryless continuous distribution.

To use memorylessness to recompute $\mathbb{E}[X]$, we need to split an expectation according to which of several events occurred. This is the expectation version of the law of total probability from Lecture 2.

Theorem 4.8 (Total expectation theorem)

Let $A_1, \dots, A_n$ be a partition of Ω with $\mathbb{P}(A_i) > 0$ for each i , and define the conditional expectation given an event, $\mathbb{E}[X \mid A_i] = \sum_x x \mathbb{P}(X = x \mid A_i)$. Then

$$\mathbb{E}[X] = \sum_{i=1}^n \mathbb{P}(A_i) \mathbb{E}[X \mid A_i]. \quad (4.55)$$

Proof. For each value x , the law of total probability gives $\mathbb{P}(X = x) = \sum_i \mathbb{P}(A_i) \mathbb{P}(X = x \mid A_i)$. Substituting into the definition of expectation and swapping the order of summation (justified by absolute convergence),

$$\begin{aligned} \mathbb{E}[X] &= \sum_x x \mathbb{P}(X = x) = \sum_x x \sum_i \mathbb{P}(A_i) \mathbb{P}(X = x \mid A_i) \\ &= \sum_i \mathbb{P}(A_i) \sum_x x \mathbb{P}(X = x \mid A_i) = \sum_i \mathbb{P}(A_i) \mathbb{E}[X \mid A_i]. \end{aligned} \quad \square$$

Example 4.2 (Geometric mean via memorylessness)

Let $X \sim \text{Geometric}(p)$. Condition on the outcome of the first trial, using the partition $\{X = 1\}$ (first trial succeeds) and $\{X > 1\}$ (first trial fails):

$$\mathbb{E}[X] = \mathbb{P}(X = 1) \mathbb{E}[X \mid X = 1] + \mathbb{P}(X > 1) \mathbb{E}[X \mid X > 1]. \quad (4.56)$$

- $\mathbb{P}(X = 1) = p$ (success on the first try), and then $X = 1$ exactly, so $\mathbb{E}[X \mid X = 1] = 1$.
- $\mathbb{P}(X > 1) = 1 - p$. If the first trial failed, the geometric “tells us we start again”: by memorylessness the remaining wait $X - 1$ is again $\text{Geometric}(p)$, and we have already

spent one trial. So

$$\mathbb{E}[X \mid X > 1] = \mathbb{E}[(X - 1) + 1 \mid X > 1] = \mathbb{E}[X] + 1. \quad (4.57)$$

Substituting,

$$\begin{aligned} \mathbb{E}[X] &= p \cdot 1 + (1 - p)(\mathbb{E}[X] + 1) \\ &= p + \mathbb{E}[X] + 1 - p\mathbb{E}[X] - p \\ &= \mathbb{E}[X] + 1 - p\mathbb{E}[X]. \end{aligned} \quad (4.58)$$

Subtracting $\mathbb{E}[X]$ from both sides gives $0 = 1 - p\mathbb{E}[X]$, i.e. $p\mathbb{E}[X] = 1$, so

$$\mathbb{E}[X] = \frac{1}{p}, \quad (4.59)$$

matching the tail sum computation.

Two checks on this argument.

1. **Finiteness.** Subtracting $\mathbb{E}[X]$ from both sides is only legitimate because $\mathbb{E}[X] < \infty$, which we already know from the tail sum. (If $\mathbb{E}[X] = \infty$, the equation $\infty = \infty + 1 - \infty$ says nothing.)
2. **Sanity check the answer.** Any correct formula must satisfy $\mathbb{E}[X] \geq 1$ (you need at least one trial) and $\mathbb{E}[X] \rightarrow \infty$ as $p \rightarrow 0$. A sign slip in the algebra can produce $\mathbb{E}[X] = \frac{1}{1+p}$, which fails both tests, since it is less than 1 for every $p > 0$.

Theorem 4.9 (Geometric variance (supplementary))

If $X \sim \text{Geometric}(p)$, then

$$\text{Var}(X) = \frac{1 - p}{p^2}. \quad (4.60)$$

Proof. We use the same first-trial conditioning, now on $\mathbb{E}[X^2]$ (which is finite, since the tail of X decays geometrically).

$$\mathbb{E}[X^2] = \mathbb{P}(X = 1)\mathbb{E}[X^2 \mid X = 1] + \mathbb{P}(X > 1)\mathbb{E}[X^2 \mid X > 1]. \quad (4.61)$$

- Given $X = 1$: $X^2 = 1$, so $\mathbb{E}[X^2 \mid X = 1] = 1$.
- Given $X > 1$: by memorylessness, $X = 1 + X'$ where $X' \sim \text{Geometric}(p)$ has the same distribution as X . Expanding the square,

$$\mathbb{E}[X^2 \mid X > 1] = \mathbb{E}[(1 + X')^2] = 1 + 2\mathbb{E}[X'] + \mathbb{E}[X'^2] = 1 + 2\mathbb{E}[X] + \mathbb{E}[X^2]. \quad (4.62)$$

Substituting, with $\mathbb{E}[X] = 1/p$:

$$\begin{aligned} \mathbb{E}[X^2] &= p \cdot 1 + (1 - p)(1 + 2\mathbb{E}[X] + \mathbb{E}[X^2]) \\ &= p + (1 - p) + 2(1 - p)\mathbb{E}[X] + (1 - p)\mathbb{E}[X^2] \\ &= 1 + \frac{2(1 - p)}{p} + (1 - p)\mathbb{E}[X^2]. \end{aligned} \quad (4.63)$$

Move the $\mathbb{E}[X^2]$ terms to the left:

$$\mathbb{E}[X^2] - (1-p)\mathbb{E}[X^2] = p\mathbb{E}[X^2] = 1 + \frac{2(1-p)}{p} = \frac{p+2-2p}{p} = \frac{2-p}{p}. \quad (4.64)$$

Dividing by p , $\mathbb{E}[X^2] = \frac{2-p}{p^2}$. Finally,

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}. \quad (4.65)$$

□

Cross-check with the factorial moment trick. Differentiating $\sum_{k \geq 0} x^k = \frac{1}{1-x}$ twice gives $\sum_{k \geq 2} k(k-1)x^{k-2} = \frac{2}{(1-x)^3}$. With $x = 1-p$,

$$\mathbb{E}[X(X-1)] = \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-1}p = p(1-p) \sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} = p(1-p) \cdot \frac{2}{p^3} = \frac{2(1-p)}{p^2}, \quad (4.66)$$

so $\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] = \frac{2(1-p)}{p^2} + \frac{p}{p^2} = \frac{2-p}{p^2}$, in agreement.

4.6 Poisson Merging and Splitting

Poisson random variables model counts of arrivals (packets at a router, calls at a switchboard, customers at a queue). Two structural facts make them especially convenient: independent Poisson counts *merge* into a Poisson count, and randomly *splitting* a Poisson count produces independent Poisson counts.

Theorem 4.10 (Poisson merging (sum of independent Poissons))

Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent. Then

$$X + Y \sim \text{Poisson}(\lambda + \mu), \quad (4.67)$$

with mean $\mathbb{E}[X + Y] = \lambda + \mu$.

Proof (supplementary). Fix $n \in \{0, 1, 2, \dots\}$. Since $X, Y \geq 0$, the event $\{X + Y = n\}$ splits into

the disjoint events $\{X = k, Y = n - k\}$ for $k = 0, \dots, n$:

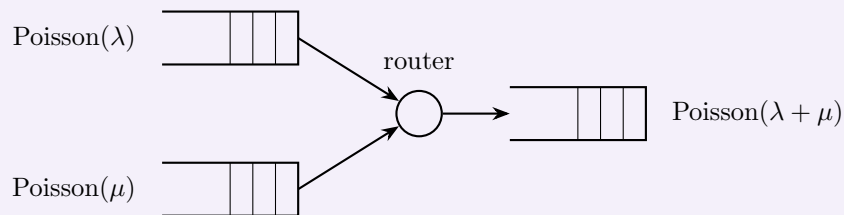
$$\begin{aligned}
 \mathbb{P}(X + Y = n) &= \sum_{k=0}^n \mathbb{P}(X = k, Y = n - k) \\
 &= \sum_{k=0}^n \mathbb{P}(X = k) \mathbb{P}(Y = n - k) && \text{(independence)} \\
 &= \sum_{k=0}^n \frac{e^{-\lambda} \lambda^k}{k!} \cdot \frac{e^{-\mu} \mu^{n-k}}{(n-k)!} && \text{(Poisson PMFs)} \\
 &= e^{-(\lambda+\mu)} \sum_{k=0}^n \frac{\lambda^k \mu^{n-k}}{k! (n-k)!} && \text{(collect exponentials)} \\
 &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^n \frac{n!}{k! (n-k)!} \lambda^k \mu^{n-k} && \text{(multiply and divide by } n!) \\
 &= \frac{e^{-(\lambda+\mu)}}{n!} \sum_{k=0}^n \binom{n}{k} \lambda^k \mu^{n-k} \\
 &= \frac{e^{-(\lambda+\mu)} (\lambda + \mu)^n}{n!}. && \text{(binomial theorem)} \tag{4.68}
 \end{aligned}$$

This is exactly the $\text{Poisson}(\lambda + \mu)$ PMF evaluated at n . The mean $\lambda + \mu$ also follows directly from linearity of expectation. $\square$

Corollary 4.4 (Many independent Poissons)

If $X_1, \dots, X_m$ are independent with $X_i \sim \text{Poisson}(\lambda_i)$, then $\sum_{i=1}^m X_i \sim \text{Poisson}(\sum_{i=1}^m \lambda_i)$. (Apply the theorem repeatedly; $X_1 + \dots + X_k$ is independent of X_{k+1} .)

Networking picture: merging. Two independent packet streams feed a router through separate buffers. If the numbers of packets arriving in each buffer during an interval are $\text{Poisson}(\lambda)$ and $\text{Poisson}(\mu)$, the number of packets in the merged output buffer is $\text{Poisson}(\lambda + \mu)$.



Splitting is the reverse operation: each arriving packet is routed to one of two outputs by an independent coin flip. Given the total count $X = n$, the number Y sent to the first output is the number of successes in n independent Bernoulli(p) trials. This is a “binomial with a random number of trials.”

Theorem 4.11 (Poisson splitting)

Let $X \sim \text{Poisson}(\lambda)$, and suppose that, conditioned on $X = n$, the random variable Y satisfies

$$Y \mid X = n \sim \text{Binomial}(n, p). \quad (4.69)$$

Let $Z = X - Y$ be the remainder. Then

$$Y \sim \text{Poisson}(\lambda p), \quad Z \sim \text{Poisson}(\lambda(1 - p)), \quad \text{and} \quad Y \perp Z. \quad (4.70)$$

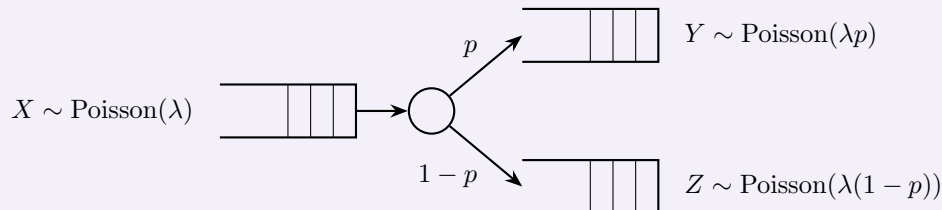
Proof (supplementary). We compute the joint PMF of (Y, Z) and show that it factors. Fix integers $a, b \geq 0$. The event $\{Y = a, Z = b\}$ is the same as $\{X = a + b, Y = a\}$ (the total is $a + b$, of which a went to the first output). By the chain rule,

$$\begin{aligned} \mathbb{P}(Y = a, Z = b) &= \mathbb{P}(X = a + b) \mathbb{P}(Y = a \mid X = a + b) \\ &= \frac{e^{-\lambda} \lambda^{a+b}}{(a+b)!} \cdot \binom{a+b}{a} p^a (1-p)^b && \text{(Poisson, then binomial)} \\ &= \frac{e^{-\lambda} \lambda^{a+b}}{(a+b)!} \cdot \frac{(a+b)!}{a! b!} p^a (1-p)^b && \text{(expand the binomial coefficient)} \\ &= e^{-\lambda} \frac{\lambda^a p^a}{a!} \cdot \frac{\lambda^b (1-p)^b}{b!} && \text{(cancel } (a+b)!, \text{ split } \lambda^{a+b}) \\ &= \underbrace{\frac{e^{-\lambda p} (\lambda p)^a}{a!}}_{\text{Poisson}(\lambda p) \text{ PMF at } a} \cdot \underbrace{\frac{e^{-\lambda(1-p)} (\lambda(1-p))^b}{b!}}_{\text{Poisson}(\lambda(1-p)) \text{ PMF at } b}, && \text{(split } e^{-\lambda} = e^{-\lambda p} e^{-\lambda(1-p)}) \end{aligned} \quad (4.71)$$

Marginal of Y . Sum over $b \geq 0$. The second factor is a PMF in b , so it sums to 1, leaving $\mathbb{P}(Y = a) = \frac{e^{-\lambda p} (\lambda p)^a}{a!}$. Thus $Y \sim \text{Poisson}(\lambda p)$. Summing over a instead shows $Z \sim \text{Poisson}(\lambda(1 - p))$.

Independence. The joint PMF equals the product of the two marginals for every (a, b) , which is the definition of $Y \perp Z$. $\square$

Why the independence is surprising. If the total X were a fixed number n , then Y and $Z = n - Y$ would be perfectly (negatively) dependent: learning Y would tell you Z exactly. It is the *randomness of the total count* that makes the two outputs independent. A large value of Y suggests X was large, which exactly offsets what it would otherwise say about Z . This is special to the Poisson.



4.7 Conditional PMFs and Conditional Expectation

Definition 4.4 (Conditional PMF)

Let X, Y be discrete random variables and let y satisfy $p_Y(y) > 0$. The **conditional PMF of X given $Y = y$** is

$$p_{X|Y}(x | y) \triangleq \mathbb{P}(X = x | Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}. \quad (4.72)$$

It is a genuine PMF in x . For each fixed y with $p_Y(y) > 0$:

- $p_{X|Y}(x | y) \geq 0$, since it is a ratio of a probability and a positive number;
- it sums to one over x , by marginalizing the joint PMF:

$$\sum_x p_{X|Y}(x | y) = \frac{1}{p_Y(y)} \sum_x p_{X,Y}(x, y) = \frac{p_Y(y)}{p_Y(y)} = 1. \quad (4.73)$$

So for each fixed value y , “ X given $Y = y$ ” is just another random variable with its own distribution, and we can take its expectation. (Note that the fraction is $p_{X,Y}(x, y)/p_Y(y)$, a ratio of PMF values, rather than $\mathbb{P}(X \cap Y)/\mathbb{P}(Y)$, since X and Y are random variables rather than events.)

Definition 4.5 (Conditional expectation given $Y = y$)

For y with $p_Y(y) > 0$, the **conditional expectation of X given $Y = y$** is

$$\mathbb{E}[X | Y = y] \triangleq \sum_x x \cdot p_{X|Y}(x | y) = \sum_x x \cdot \mathbb{P}(X = x | Y = y). \quad (4.74)$$

Like any expectation, $\mathbb{E}[X | Y = y]$ is just a **real number**.

By “sweeping” through the values of y in (4.74), we get a (possibly different) number $\mathbb{E}[X | Y = y]$ for every value y . That is, $y \mapsto \mathbb{E}[X | Y = y]$ is a *function* of y .

Definition 4.6 (Conditional expectation as a random variable)

Define $g(y) = \mathbb{E}[X | Y = y]$. The **conditional expectation of X given Y** is the random variable

$$\mathbb{E}[X | Y] \triangleq g(Y). \quad (4.75)$$

It is a function of Y , and therefore $\mathbb{E}[X | Y]$ is itself a **random variable**: it takes the value $g(y)$ with probability $p_Y(y)$.

Notation: $\mathbb{E}[X | Y = y]$ **versus** $\mathbb{E}[X | Y]$. These are different objects. $\mathbb{E}[X | Y = y]$ (lowercase y , a specific value) is a number. $\mathbb{E}[X | Y]$ (capital Y) is a random variable.

Example 4.3 (Die roll, conditioned on parity)

Roll a fair six-sided die. Let X be the result and $Y = \mathbf{1}\{X \text{ is even}\}$.

Given $Y = 1$, X is uniform on $\{2, 4, 6\}$, so $\mathbb{E}[X | Y = 1] = \frac{2+4+6}{3} = 4$. Given $Y = 0$, X is uniform on $\{1, 3, 5\}$, so $\mathbb{E}[X | Y = 0] = \frac{1+3+5}{3} = 3$. Hence

$$\mathbb{E}[X | Y] = g(Y) = \begin{cases} 4 & \text{if } Y = 1, \\ 3 & \text{if } Y = 0, \end{cases} \quad \text{i.e.} \quad \mathbb{E}[X | Y] = 3 + Y. \quad (4.76)$$

This is a random variable taking the values 3 and 4, each with probability $\frac{1}{2}$. Notice that its mean is $\frac{1}{2}(3) + \frac{1}{2}(4) = 3.5 = \mathbb{E}[X]$. That is no coincidence, as the next lecture will show.

Law of Iterated Expectation (Tower Rule): TBD next lecture. This was on today's agenda but was not reached in lecture. It states that $\mathbb{E}[\mathbb{E}[X | Y]] = \mathbb{E}[X]$; the total expectation theorem of Section 4.5 is a special case.

4.8 Summary: Common Discrete Distributions

Distribution	PMF $\mathbb{P}(X = k)$	Support	$\mathbb{E}[X]$	$\text{Var}(X)$
Bernoulli(p)	$p^k(1-p)^{1-k}$	$\{0, 1\}$	p	$p(1-p)$
Binomial(n, p)	$\binom{n}{k}p^k(1-p)^{n-k}$	$\{0, \dots, n\}$	np	$np(1-p)$
Geometric(p)	$(1-p)^{k-1}p$	$\{1, 2, \dots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$	$\{0, 1, 2, \dots\}$	λ	λ

End of Lecture 4.