# 11 Lecture 11, November 5, 2014

## 11.1 Improvement to the proof of Shannon's coding theorem, using linear codes

Very commonly, in Algorithms, we have a tradeoff between how much randomness we use, and efficiency.

But sometimes we can actually improve our efficiency by carefully eliminating some of the randomness we're using. Roughly, the intuition is that some of the randomness is going not toward circumventing a barrier (especially, leaving the adversary in the dark about what we are going to do), but just into noise.[8]

A case in point is the proof of Shannon's Coding Theorem. In a previous lecture we proved the theorem as follows: we first built an encoding map $\mathcal{E} : \{0,1\}^k \to \{0,1\}^n$ by sampling a uniformly random function; then, we had to delete up to half the codewords to eliminate all kinds of fluctuations in which codewords fell too close to one another.

It turns out that this messy solution can be avoided. The key observation is that our analysis depended only on pairwise data about the code—basically, pairwise distances between codewords. "Higher level" structure (mutual distances among triples, etc.) didn't feature in the analysis. So the argument will still go through with a pairwise-independently constructed code. So we'll do this now, and in the process we'll see how this helps.

Sample $\mathcal{E}$ from the following *pairwise independent* family of functions $\{0,1\}^k \to \{0,1\}^n$. Select $k$ vectors $v_1, \ldots, v_k$ iid $\in_u \{0,1\}^n$. Now map the vector $(x_1, \ldots, x_k)$ to $\sum_1^k x_i v_i$. This is, of course, a linear map:

$$(\text{ message } x) \begin{pmatrix} --- & v_1 & --- \\ --- & v_2 & --- \\ --- & \cdots & --- \\ --- & v_k & --- \end{pmatrix} = (\text{ codeword })$$

The message $\bar{0} \in \{0,1\}^k$ is always mapped to the codeword $\bar{0} \in \{0,1\}^n$, and every other codeword is uniformly distributed in $\{0,1\}^n$. It is not hard to see that the images of messages are pairwise independent. (Including even the image of the $\bar{0}$ message.) Looking back at the analysis of the error probability in section (8.1), it had two parts, in each of which we bounded the probability of one of the following two sources of error:

Bad$_1$: $H(\mathcal{E}(X) + R, \mathcal{E}(X)) \geq (p+\delta)n$. That is to say, the error vector $R$ has weight (number of 1's) at least $(p+\delta)n$. This analysis is of course unchanged, and is independent of the choice of the code. As before the bound is $\leq 2^{-D_2(p+\delta\|p)n}$.

Bad$_2$: $\exists X' \neq X : H(\mathcal{E}(X) + R, \mathcal{E}(X')) \leq (p+\delta)n$. For this, pairwise independence is enough to obtain an analysis similar to before. Specifically, for any pair $X \neq X'$ and any $R$, the rv (which now depends only on the choice of code) $\mathcal{E}(X) + R - \mathcal{E}(X')$ is uniformly distributed in $\{0,1\}^n$ so, by the same union bound as before, we can bound the probability of this error event by $2^{k-n(D_2(p\|1/2)-\varepsilon/2)} = 2^{-n\varepsilon/2}$.

So, the analysis is unchanged from before insofar as bounding $\Pr_{\mathcal{E},X,R}(\text{Error})$ above by $2^{1-cn}$ for some $c > 0$ that depends only on $p, \varepsilon$. (See Eqn. (21).)

Next, just as before, we wish to remove $\mathcal{E}$ and $X$ from the randomization in the analysis. The first step is just as before: there exists a *specific* code $\mathcal{E}$ achieving $\Pr_{X,R}(\text{Error}|\mathcal{E}) \leq 2^{1-cn}$. And if we wish to find it slightly more quickly by trying codes at random, we can easily do so just by settling for a bound of $2^{2-cn}$. (On average we only need to try two $\mathcal{E}$'s, but it still takes a while to calculate or at least estimate with high probability the $\Pr_{X,R}(\text{Error}|\mathcal{E})$ for each $\mathcal{E}$ we try. The key computational difficulty is actually in finding the closest codeword to the randomly generated $\mathcal{E}(X) + R$; this is a hard problem, for randomly generated linear codes. So, in practice, we use much more structured linear codes.)

[8] *If you carry overnight gear, you'll certainly spend the night on the mountain* – a climbing instructor I knew

Finally, what changes comes next. Rewrite $\Pr_{X,R}(\text{Error}|\mathcal{E}) = E_X \Pr_R(\text{Error}|\mathcal{E}, X)$. The interesting thing is that we may take the code to have the property that for all $X_1, X_2$, $\Pr_R(\text{Error}|\mathcal{E}, X_1) = \Pr_R(\text{Error}|\mathcal{E}, X_2)$. This is because something even stronger is true: once you fix $\mathcal{E}$, there is an optimal decoder $\mathcal{D}$ (maximizing probability of decoding, w.r.t. uniformly chosen messages), such that exactly the same noise vectors $R$ cause errors for $X_1$ as for $X_2$. Specifically, suppose that $\mathcal{D}(\mathcal{E}(X_1) + R) = X'$. That means that

$$X' \in \text{argmin}_Z \, H(R + \mathcal{E}(X_1) - \mathcal{E}(Z)). \tag{30}$$

Here $H$ denotes the Hamming weight of the vector, i.e., the number of 1's in it; equivalently, the distance to $\bar{0}$. (Also note, $-\mathcal{E}(X)$ is the same as $+\mathcal{E}(X)$, since these are bits, but this same derivation is useful also in situations where these aren't bits, so we may as well carry the signs around.)

Eqn. 30 does not fully specify a decoding rule because there may be ties. However, in that case, since the potential $Z$'s occur with equal probabilities, we may choose any of those in the argmin without changing $\Pr_{X,R}(\text{Error}|\mathcal{E})$. So in particular we may use a decoder that, besides being minimum-distance, also commutes with translation by a codeword, that is to say, for any $X, Y$ we have

$$\mathcal{D}(Y + \mathcal{E}(X)) = \mathcal{D}(Y) + X. \tag{31}$$

Now suppose that for a specific message $X$, noise $R$ causes error. That is to say, $\mathcal{D}(\mathcal{E}(X) + R) \neq X$. Then for the message $\bar{0}$, using Eqn. 31, $\mathcal{D}(\mathcal{E}(\bar{0}) + R) = -X + \mathcal{D}(\mathcal{E}(X) + R) \neq \bar{0}$. So error event on $(X, R)$ is a property of $R$ alone and does not depend on $X$. The probability of error is therefore the same for all $X$.

So, in our linear code $\mathcal{E}$, the error probabilities associated with all the messages are equal to the overall error probability of $\mathcal{E}$, i.e., $\Pr_{X,R}(\text{Error}|\mathcal{E})$. There is no need to throw away any $X$'s with high error probabilities.

## 11.2 Variance and the Chebyshev inequality

Let $X$ be a real-valued rv. If $E(X)$ and $E(X^2)$ are both well-defined and finite, let $\text{Var}(X) = E((X - E(X))^2)$. Expanding and applying linearity of expectation, this is also $= E(X^2) - E(X)^2$.

Note that If $c \in \mathbb{R}$ then since the variance is homogenous and quadratic, $\text{Var}(cX) = c^2 \text{Var}(X)$.

**Lemma 46 (Chebyshev)** *If $E(X) = \theta$, then $\Pr(|X - \theta| > \lambda\sqrt{\text{Var}(X)}) < 1/\lambda^2$.*

**Proof:** $\Pr\left(|X - \theta| > \lambda\sqrt{\text{Var}(X)}\right) = \Pr\left((X - \theta)^2 > \lambda^2 \text{Var}(X)\right) < 1/\lambda^2$ by the Markov inequality (Lem. 10). □

A frequently useful corollary is:

**Corollary 47** *Suppose $X$ is a nonnegative rv. Then $\Pr(X = 0) \leq \frac{\text{Var}(X)}{(E(X))^2}$.*

## 11.3 Pairwise independence and the second-moment inequality

A common situation in which we use Chebyshev's inequality is when we have many variables which are not fully independent, but are pairwise independent (or nearly so).

**Definition 48 (Pairwise and $k$-wise independence)** *A set of rvs are pairwise independent if every pair of them are independent; this is a weaker requirement than that all be independent. Likewise, the variables are $k$-wise independent if every subset of size $k$ is independent.*

**Definition 49 (Covariance)** *The covariance of two real-valued rvs $X, Y$ is (if well-defined)* $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$.

*Exercise:* Show that if $X$ and $Y$ are independent then $\mathrm{Cov}(X, Y) = 0$, but that the converse need not be true.

*Exercise:* If $X = \sum_1^n X_i$, $\mathrm{Var}\, X = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$.

**Corollary 50** *If $X_1, \ldots, X_n$ are pairwise independent real rvs with well-defined variances, then $\mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i)$. If $\overline{X} = \frac{1}{n} \sum X_i$, then $E(\overline{X}) = E(X_1)$ and $\mathrm{Var}(\overline{X}) = \frac{1}{n} \mathrm{Var}(X_1)$.*

*Exercise: Apply the Chebyshev inequality to obtain:*

**Lemma 51 (2nd moment inequality)** *If $X_1, \ldots, X_n$ are identically distributed, pairwise-independent real rvs with finite 1st and 2nd moments then $P(|\overline{X} - E(\overline{X})| > \lambda \sqrt{\frac{\mathrm{Var}(X_1)}{n}}) < 1/\lambda^2$.*

**Corollary 52 (Weak Law)** *Pairwise independent rvs obey the weak law of large numbers. Specifically, if $X_1, \ldots, X_n$ are identically distributed, pairwise-independent real rvs with finite variance then for any $\varepsilon$, $\lim_{n \to \infty} P(|\overline{X} - E(\overline{X})| > \varepsilon) \to 0$.*

So we see that the weak law holds under a much weaker condition than full independence. When we talk about the *cardinality* of sample spaces, we'll see why pairwise (or small $k$-wise) independence has a huge advantage over full independence, so that it is often desirable in computational settings to make do with limited independence.

## 11.4 Threshold for $K_4$ in $G(n, p)$

Working with low moments of random variables can be incredibly effective, even when we are not specifically looking for limited-independence sample spaces. Here is a prototypical example. "When" does a 4-clique (written $K_4$) show up in a random graph selected from the distribution $G(n, p)$? We have in mind that we are "turning the knob" on $p$. When $p = 0$, of course with probability 1 there is no subgraph isomorphic to $K_4$. When $p = 1$, with probability 1 there is such a subgraph, in fact, $\binom{n}{4}$ of them. In between, for any finite $n$, the probability is finite. But we won't take $n$ finite, we will take it tending to $\infty$.

So the question is,[9] can we identify a function $\pi(n)$ such that in the model $G(n, p(n))$, with $[\![K_4]\!]$ denoting the event that there is a $K_4$ in the random graph $G$,

(a) If $p(n) \in o(\pi(n))$, then $\lim_n \Pr([\![K_4]\!]) = 0$.

(b) If $p(n) \in \omega(\pi(n))$, then $\lim_n \Pr([\![K_4]\!]) = 1$.

Such a function $\pi(n)$ is known as the *threshold* for appearance of $K_4$. It turns out that under weak assumptions, monotone events—events that hold in $G'$ if they hold in $G$ and $G \subseteq G'$—always have such threshold functions. In particular this is true for any monotone graph property, that is, a property that is preserved under permutation of the vertices. (The statement is even stronger: for any $\varepsilon > 0$ there is a $p(n)$ such that $\Pr_{p(n)}(\text{property}) \leq \varepsilon$ and $\Pr_{p(n)+O(1/\log n)}(\text{property}) \geq 1 - \varepsilon$. See [26].)

Let $S \subset \{1, \ldots, n\}$, $|S| = 4$. Let $X_S$ be the event that $K_4$ occurs as a subgraph of $G$ at $S$—that is, when you look at those four vertices, all the edges between them are present. [10] Conflating $X_S$ with its indicator function and letting $X$ be the number of $K_4$'s in $G$, we have

$$X = \sum_S X_S$$

[9] Recall $p(n) \in o(\pi(n))$ means that $\limsup p(n)/\pi(n) = 0$, and $p(n) \in \omega(\pi(n))$ means that $\limsup \pi(n)/p(n) = 0$.

[10] More generally the method we are studying can be used to establish the probability of any fixed graph $H$ occurring as a subgraph in $G$, that is, there is an injection of the vertices of $H$ into the vertices of $G$ such that every edge of $H$ is present in $G$. This is different from asking that $H$ occur as a *induced* subgraph of $G$, in which case one also demands that the non-edges of $H$ be non-edges in $G$. That is an interesting question too but different in an essential way: the event is not monotone in $G$.

and

$$E(X) = \binom{n}{4} p^6.$$

We are interested in $\Pr(X > 0)$. Let $\pi(n) = n^{-2/3}$.

(a) For $p(n) \in o(\pi(n))$, $E(X) \in o(1)$, so $\Pr([\![K_4]\!]) \in o(1)$ and therefore $\lim_n \Pr([\![K_4]\!]) = 0$.

(b) For $1 > p(n) \in \omega(\pi(n))$, $E(X) \in \omega(1)$. We'd like to conclude that likely $X > 0$ but we do not have enough information to justify this, as it could be that $X$ is usually 0 and occasionally very large. We will exclude that possibility by studying the next moment of the distribution.

Before carrying out this calculation, though, we have to make one important note. Since the event $[\![K_4]\!]$ is monotone, $[p \leq p'] \Rightarrow [\Pr_{G(n,p)}[\![K_4]\!] \leq \Pr_{G(n,p')}[\![K_4]\!]]$. (An easy way to see this is by modifying the probabilities the edges one by one.) This means that it is enough to show that $K_4$ "shows up" slightly above $\pi$. This is useful because some of our calculations break down far above $\pi$, not because there is anything wrong with the underlying statement but because the inequalities we use are not strong enough to be useful there and a direct calculation would need to take account of further moments.

To simplify our remaining calculations, then, let $p = n^{\varepsilon - 2/3}$ for some small $\varepsilon > 0$. (So we won't prove the entire strength of the theorem but this is only to keep expressions simple; the same calculation can be reproduced closer to $\pi$, as required by the theorem statement.)

By an earlier exercise,

$$\mathrm{Var}(X) = \sum_S \mathrm{Var}(X_S) + \sum_{S \neq T} \mathrm{Cov}(X_S, X_T)$$

$X_S$ is a coin (or Bernoulli rv) with probability $p^6$ of coming up "heads". As the reader should check, the variance of such an rv is $p^6(1 - p^6)$. So

$$\mathrm{Var}(X_S) > 0 \text{ for } 0 < p < 1$$

The covariance terms are more interesting.

1. If $|S \cap T| \leq 1$, no edges are shared, so the events are independent and $\mathrm{Cov}(X_S, X_T) = 0$.
2. If $|S \cap T| = 2$, one edge is shared, and a total of 11 specific edges must be present for both cliques to be present. A simple way to bound the covariance is (since $E(X_S), E(X_T) \geq 0$) that $\mathrm{Cov}(X_S, X_T) \leq E(X_S X_T) = p^{11}$.
3. If $|S \cap T| = 3$, three edges are shared, and a specific 9 edges must be present for both cliques to be present. Similarly to the previous case, $\mathrm{Cov}(X_S, X_T) \leq p^9$.

$$\begin{aligned}
\mathrm{Var}(X) &\leq \binom{n}{4} p^6(1 - p^6) + \binom{n}{2,2,2} p^{11} + \binom{n}{3,1,1} p^9 \\
&\in O(n^4 p^6 + n^6 p^{11} + n^5 p^9) \\
&= O(n^{4+6\varepsilon-4} + n^{6+11\varepsilon-22/3} + n^{5+9\varepsilon-6}) \\
&= O(n^{6\varepsilon})
\end{aligned}$$

This gives us the key piece of information:

$$\frac{\mathrm{Var}(X)}{(E(X))^2} \in \frac{O(n^{6\varepsilon})}{\Omega((n^4 p^6)^2)} = \frac{O(n^{6\varepsilon})}{\Omega(n^{12\varepsilon})} = O(n^{-6\varepsilon}) \subseteq o(1)$$

and we have only to apply the Chebyshev inequality (Cor. 47) to conclude that $\Pr(X = 0) \in o(1)$ and so $\lim_n \Pr([\![K_4]\!]) = 1$.