10-704: Information Processing and Learning **Spring 2015**

Lecture 18: March 26

Lecturer: Aarti Singh *Scribes: Che Zheng*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

18.1 Review of Gaussian Channel

Suppose we have a single variable Gaussian channel with output Y, input X and noise Z. Also suppose X is independent of Z, i.e. $X \perp Z$, and $Z \sim N(0, \sigma^2)$.

Theorem 18.1 *Suppose $Y = X + Z$, $X, Y, Z \in \mathbb{R}$ is a Gaussian channel and $Z \sim N(0, \sigma^2)$. Given power constraint P, i.e. $\mathbb{E}[X^2] \leq P$, then*

$$Capacity = \frac{1}{2}\log(1 + \frac{P}{\sigma^2})$$

Now consider the multi-dimensional case. Suppose our input, output and noise now are in $\mathbb{R}^n$ space, and Z to is draw from $N(0, \sigma^2 I_{n\times n})$. We have independent power constraint for each sub channel as $E[X_i^2] \leq P$ for $i = 1, \ldots, n$. Then the capacity is directly n times the capacity of one channel.

Theorem 18.2 *Suppose $Y = X + Z$, $X, Y, Z \in \mathbb{R}^n$ is a multivariate Gaussian channel and $Z \sim N(0, \sigma^2 I)$. Given independent power constraint on each X_i, i.e. $E[X_i^2] \leq P$. Then the capacity of this channel is given by*

$$Capacity = \frac{n}{2}\log(1 + \frac{P}{\sigma^2})$$

Now consider the case if a global power constraint like $E[||X||^2] \leq P$ is used. With same condition as above, we can prove the capacity is maximized when we equally distribute the power constraint over all channels, i.e. $P_i = \frac{P}{n}$ and the capacity is given as follows.

Theorem 18.3 *Suppose $Y = X + Z$, $X, Y, Z \in \mathbb{R}^n$ is a multivariate Gaussian channel and $Z \sim N(0, \sigma^2 I)$. Given universal power constraint over X, i.e. $E[||X||^2] \leq P$. Then the capacity of this channel is given by*

$$Capacity = \frac{n}{2}\log(1 + \frac{P}{n\sigma^2})$$

18.2 Independent Gaussian Channel

In above examples, our input channels have independent noise with same variance. If our noise is still independent on each channel, but with different variance, the maximum capacity is given through a "water filling" way.

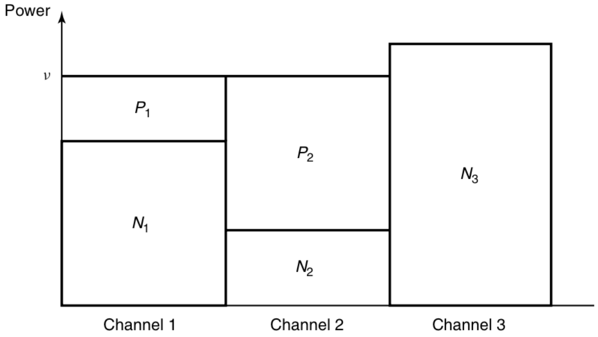

Figure 18.1: Water filling, figure from [Cover2012]

Theorem 18.4 *Suppose $Y = X+Z$, $X, Y, Z \in \mathbb{R}^n$ is a multivariate Gaussian channel, $Z \sim N(0, diag(\sigma_1^2, \dots, \sigma_n^2))$. Given universal power constraint, i.e. $E[||X||^2] \leq P$, then the capacity of this channel is given through a "water-filling" way. That is, the power allocated for each channel P_i is $(constant - \sigma_i^2)^+$, where the constant is chosen so that the total power $\sum_i P_i$ is P.*

Proof: We know capacity C is defined as

$$\begin{aligned} C =& \max_{p(x)} I(X^n, Y^n) \\ =& \max_{p(x)} (H(Y^n) - H(Y^n|X^n)) \\ =& \max_{p(x)} (H(Y^n)) - H(Z^n), \quad \text{Since } Y = X + Z \text{ and } X \perp Z \\ =& \max_{p(x)} \sum_{i=1}^{n} H(Y_i) - H(Z_i), \quad \text{since } Y_i\text{s are independent and so are } Z_i\text{s} \end{aligned}$$

We know $Y_i = X_i + Z_i$, so $E[Y_i^2] = E[(X_i + Z_i)^2] = E[X_i^2] + E[Z_i^2] = P_i + \sigma^2$. We know for a given variance, normal distribution maximize the entropy, thus $H(Y_i) \leq \frac{1}{2} \log 2\pi e(P_i + \sigma_i)$.

$$\begin{aligned} \leq & \max_{\{P_i\}_{i=1}^n} \frac{1}{2} \sum_{i=1}^{n} \log 2\pi e(P_i + \sigma_i^2) - \frac{1}{2} \sum_{i=1}^{n} \log 2\pi e \sigma_i^2 \\ = & \max_{\{P_i\}_{i=1}^n} \frac{1}{2} \sum_{i=1}^{n} \log \left(1 + \frac{P_i}{\sigma_i^2}\right) \end{aligned}$$

Since the $P_i \geq 0$ and $\sum_i P_i \leq P$, the Lagrangian multiplier of the above optimization problem is

$$\begin{aligned} \mathcal{L} =& \frac{1}{2} \sum_{i=1}^{n} \log \left(1 + \frac{P_i}{\sigma^2}\right) + \lambda(\sum_i P_i - P) + \lambda_i P_i \\ \frac{\partial \mathcal{L}}{\partial P_i} =& \frac{1}{2} \frac{1}{1 + \frac{P_i}{\sigma_i^2}} \cdot \frac{1}{\sigma_i^2} + \lambda + \lambda_i = 0 \\ \Rightarrow P_i + \sigma_i^2 =& constant, \forall i \quad \text{s.t.} \lambda_i = 0 \end{aligned}$$

Since complementary slackness implies that either $P_i = 0$ or $\lambda_i = 0$, the solution is either to put no power in a channel or to put enough power so that the sum of power and noise variance is a constant for all channels with non-zero power. Thus, we are putting more power to less noisy channels through a water filling way, where we first try to add power to least noisy channels until its "height" is same with the second least one, and continue until all power is allocated.

■

18.3 Correlated Gaussian Channel

Now we consider the case where Z is no longer independent on each channel, which means Σ_Z can be arbitrary covariance matrix. Suppose we still have the universal power constraint $E[||X||^2] \leq P$.

Theorem 18.5 *Suppose $Y = X + Z$, $X, Y, Z \in \mathbb{R}^n$ is a multivariate Gaussian channel, $Z \sim N(0, \Sigma_Z)$ and $X \perp Z$. Given universal power constraint i.e. $E[||X||^2] \leq P$, then the maximum capacity is achieved through spectral water filling.*

Proof: Consider the eigenvalue decomposition of Σ_Z into $U\Lambda U^T$, where U is normalized orthogonal matrix and Λ is a diagonal matrix. Then we can restate the problem in spectral domain as

$$\begin{aligned} Y =& X + Z \\ U^T Y =& U^T X + U^T Z \\ \bar{Y} =& \bar{X} + \bar{Z}, \bar{Z} \sim N(0, \Lambda) \end{aligned}$$

The original power constraint can be written as $tr(\Sigma_X) \leq P$ and translating this to $\bar{X}$ we have $tr(\Sigma_{\bar{X}}) \leq P$.

We know for $X \in \mathbb{R}^n \sim N(0, \Sigma_X)$, $H(X) = \frac{1}{2}\log(2\pi e)^n |\Sigma_X|$. Since $X \perp Z$, then $\Sigma_Y = \Sigma_X + \Sigma_Z$, so the capacity is

$$\begin{aligned} C =& \max_{p(x)} I(X, Y) \\ =& \max_{tr(\Sigma_X) \leq P} \frac{1}{2} \log \frac{|\Sigma_X + \Sigma_Z|}{|\Sigma_Z|}. \\ =& \max_{tr(\Sigma_X) \leq P} \frac{1}{2} \log \frac{|U^T \Sigma_X U + \Lambda|}{|\Lambda|}. \\ =& \max_{tr(\Sigma_{\bar{X}}) \leq P} \frac{1}{2} \log \frac{|\Sigma_{\bar{X}} + \Lambda|}{|\Lambda|}. \end{aligned}$$

This is maximized when $\Sigma_{\bar{X}}$ is diagonal matrix. Thus using the conclusion above, we see that the channels are independent in the spectral domain and the problem is same as the last one but in the spectral domain. Capacity is achieved when $U^T X \sim \mathcal{N}(0, diag(P_i))$, or equivalently, $X \sim \mathcal{N}(0, U diag(P_i) U^T)$. And the capacity is maximized through spectral water filling, where the power constraint P_i for each $\bar{X}_i$ is $(constant - \Lambda_{ii})$.

■

Channels with correlation between sub channels are similar to channels with feedback since n parallel channels with correlation can be viewed as n sequential transmissions through a channel with memory. Thus, the above expression also characterizes the capacity of channels with memory (but without feedback).

It can be shown that feedback (knowledge of past Y_is at the sender) does not help increase the capacity of memoryless channel, but for channels with memory, the capacity of channel with feedback can be larger than the capacity of channel without feedback. For channels with memory, with feedback we have:

$$C_{FB} = \max_{tr(\Sigma_X)\leq P} \frac{1}{2} \log \frac{|\Sigma_{X+Z}|}{|\Sigma_Z|}$$

which can be larger than the expression for channels with memory without feedback - the difference being $|\Sigma_{X+Z}|$ instead of $|\Sigma_X + \Sigma_Z|$ in the numerator. However, the capacity increase can be bounded as

$$C_{FB} \leq \min(2C, C + \frac{1}{2})$$

where C is the capacity without feedback. For details, see [Cover2012] Sec 9.6.

18.4 Multi-Antenna Gaussian Channels

Now suppose the channel performs a linear transformation or projection $A \in \mathbb{R}^{m\times n}$ on X, which means the channel now is

$$Y = AX + Z, X \perp Z, Z \sim N(0, \sigma^2 I)$$

A real world case of these kind of channels is the multiple antennas channel in wireless communication where the receiver has m antennas and the sender has n antennas. The projection A, known as the channel matrix, may be deterministic or random. Another case is random projections used in CS and machine learning.

We first analyze the deterministic case, where A is fixed and known. Suppose the SVD decomposition of A is $U\Sigma V^T$, and the power constraint for X is still $E[||X||^2] \leq P$.

$$\begin{aligned} Y =& AX + Z \\ U^T Y =& \Sigma V^T X + U^T Z \\ \bar{Y} =& \Sigma \bar{X} + \bar{Z} \end{aligned}$$

Since U and V are orthonormal matrices, thus $E[||X||^2] = E[||\bar{X}||^2] \leq P$ and $\Sigma_{\bar{Z}} = \sigma^2 I$. Now we get multiple independent sub channels where instead of different noise variance, the sub-channels have different signal gains. Thus we still choose variance through water filling in the spectral domain, which means now we require the power constraint P_i for $\bar{X}_i$ to follow $P_i + \frac{1}{\lambda_i}\sigma^2 = constant$, where λ_i is the square of the singular value of A. The maximum capacity is given by

$$C = \max_{tr(\Sigma_{\bar{X}})\leq P} \frac{1}{2} \log \frac{|\Sigma_{\bar{X}} + \Sigma_{\bar{Z}}|}{|\Sigma_{\bar{Z}}|}$$

Since $\Sigma_{\bar{X}} = U diag(\lambda_j P_j) U^T$ and $\Sigma_{\bar{Z}} = \sigma^2 I$

$$\begin{aligned} =& \frac{1}{2} \log |I + U diag(\frac{\lambda_j P_j}{\sigma^2}) U^T| \\ =& \frac{1}{2} \log |I + diag(\frac{\lambda_j P_j}{\sigma^2})| \\ =& \frac{1}{2} \sum_{j=1}^{n} \log \left(1 + \frac{\lambda_j P_j}{\sigma^2}\right) \end{aligned}$$

Now we consider the case when A is a random matrix independent of X and Z. We will derive an upper bound on the capacity. The maximum capacity is given by

$$\begin{aligned} C &= \sup_{p(X)} I(X;Y) \\ &\leq \sup_{p(X)} I(X;Y,A) \\ &= \sup_{p(X)} \mathbb{E}[\log \frac{P(X,Y,A)}{P(X)P(Y,A)}] \\ &= \sup_{p(X)} \mathbb{E}[\log \frac{P(Y,A|X)}{P(Y,A)}] \\ &= \sup_{p(X)} \mathbb{E}[\log \frac{P(Y|A,X)P(A|X)}{P(Y|A)P(A)}] \end{aligned}$$

Since $A \perp X$, $P(A|X) = P(A)$

$$\begin{aligned} &= \sup_{p(X)} \mathbb{E}\left[\log \frac{P(X,Y|A)}{P(X|A)P(Y|A)}\right] \\ &= \sup_{p(X)} \mathbb{E}_A[I(X;Y|A)] \end{aligned}$$

For a fixed A, we use the previous result and upper bound it using the trivial bound $P_j \leq P$ - this is pretty loose, but will suffice for our purposes.

$$\leq \frac{1}{2}\mathbb{E}_A[\log |Udiag(I + \frac{\lambda_j P}{\sigma^2})U^T|]$$

Using Jensen's inequality and concavity of log det

$$\begin{aligned} &\leq \frac{1}{2}\log |\mathbb{E}_A[Udiag(I + \frac{\lambda_j P}{\sigma^2})U^T]| \\ &\leq \frac{1}{2}\log |I + \frac{P}{\sigma^2}\mathbb{E}[U\Sigma^2 U^T]| \\ &= \frac{1}{2}\log |I + \frac{P}{\sigma^2}\mathbb{E}[AA^T]| \end{aligned}$$

Suppose A_{ij} is drawn from $\mathcal{N}(0, \frac{1}{n})$ i.i.d., which is often the case in random projections, then $\mathbb{E}[AA^T] = I$. We get

$$\begin{aligned} C &\leq \frac{1}{2}\log |(1 + \frac{P}{\sigma^2})I| \\ &= \frac{m}{2}\log(1 + \frac{P}{\sigma^2}) \end{aligned}$$

Thus we have $\sup_{p(X)} I(X,Y) \sim O(m)$, which means the maximum average information between X and Y, $\sup_{p(X)} \frac{I(X,Y)}{n} = O(\frac{m}{n})$. Basically it means the average leakage of information from X to Y is limited by m/n which is typically decaying as n increases since in many applications the number of random projections needed $m \ll n$. This can be viewed as an **average privacy guarantee via random projections**. We will talk more about this in next class.

References

[Cover2012] Cover Thomas, "Elements of Information Theory"