10-704: Information Processing and Learning **Fall 2016**

Lecture 19: Nov 7

Lecturer: Aarti Singh

Note: *These notes are based on scribed notes from Spring15 offering of this course. 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.*

19.1 Review - Capacity of continuous channels

Example (Single Gaussian Channel) $Y = X + Z$, $Z \sim \mathcal{N}(0, \sigma^2)$, power constraint $E[X^2] \leq P$.

$$C = \frac{1}{2} \log \left(\frac{\sigma^2 + P}{\sigma^2} \right)$$

Example (Parallel Independent Gaussian Channels) $Y = X + Z$, $Z \sim N(0, \sigma^2 I_{n \times n})$.

- **Individual power constraint, same noise variance.** $E[X_i^2] \leq P$ for $i = 1, \ldots, n$.

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

- **Global power constraint, same noise variance.** $E[||X||^2] \leq P$.

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

  Achieved with equal power division amongst all channels $P_i = P/n$.

- **Global power constraint, different noise variance.** $Z \sim N(0, diag(\sigma_1^2, \ldots, \sigma_n^2))$, $E[||X||^2] \leq P$.

$$C = \max_{\{P_i\}_{i=1}^n} \frac{1}{2} \sum_{i=1}^{n} \log \left(1 + \frac{P_i}{\sigma_i^2} \right)$$

  and the max is achieved when P_i is $(constant - \sigma_i^2)_+$, where the constant is chosen so that the total power $\sum_i P_i$ is P, i.e. a "water-filling" solution (P_i is either 0 or $P_i + \sigma_i^2$ is a constant).

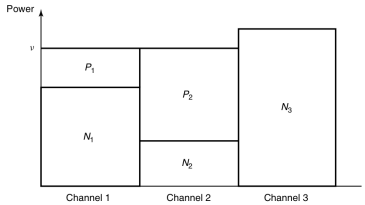

Figure 19.1: Water filling, figure from [Cover2012]

19.2 Correlated Gaussian Channel

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

Theorem 19.1 *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 global 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

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

Here $\Sigma_{\bar{X}} = U^\top \Sigma_X U$, and hence the original power constraint can be written as $tr(\Sigma_X) \leq P$ and translating this to $\bar{X}$ we have $tr(\Sigma_{\bar{X}}) = tr(U^\top \Sigma_X U) = tr(\Sigma_X U U^\top) = tr(\Sigma_X) \leq P$.

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.

19.3 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.

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 in the spectral domain 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 will 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. To see this, notice that maximum capacity is given by

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

Since Σ is diagonal matrix, the capacity is maximized for $\Sigma_{\bar{X}} = diag(P_j)$ and if λ_i is the square of the singular value of A, then we have

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

This is maximized if P_j is either 0 or $P_j + \sigma^2/\lambda_j$ is constant.

19.4 Privacy

We now consider the use of channel capacity and rate-distortion to guarantee privacy. Privacy is becoming a key concern as large datasets become publicly available. The goal of privacy is to enable releasing (a

perturbed version of) data that preserves privacy of individual data points while ensuring some utility of the privatized data for a given objective, e.g. the private data can be a perturbation of the original data that still enables prediction of the label of a new data point to some accuracy while not revealing any of the original data points. Clearly, the goals of privacy and utility are conflicting (one can always achieve perfect privacy by mapping all data points to a constant value, but such a private data has no utility) and hence it is of interest to understand the tradeoffs between the two goals.

19.4.1 Privacy via Noisy Random projections

One popular way to privatize the data is to release only a random projection of the original data points, possibly corrupted with additive Gaussian noise. In this case, if the original data is X, the privatized data is $Y = AX + Z$, where $A \in \mathbb{R}^{m\times n}$ is a random matrix independent of X and Z. One natural notion of privacy is the mutual information between the privatized data Y and the original data X, over all possible input distributions $p(X)$, i.e. precisely the capacity of the Gaussian channel linking X and Y. We will derive an upper bound on the capacity. The capacity is given by

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

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

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

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

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

where last two steps follow since $det(AB) = det(A)det(B)$, $U^\top U = I$ and $det(U^\top U) = 1$.

$$\begin{aligned} = &\frac{1}{2}\mathbb{E}_A[\log|I + \frac{P}{\sigma^2}U\Sigma^2 U^\top|] \\ = &\frac{1}{2}\mathbb{E}_A[\log|I + \frac{P}{\sigma^2}AA^\top|] \end{aligned}$$

Using Jensen's inequality and concavity of log det

$$= \frac{1}{2}\log|I + \frac{P}{\sigma^2}\mathbb{E}[AA^T]|$$

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

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

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**.

$$\sup_{p(x)} \frac{I(X,Y)}{n} \leq \frac{m}{2n} \log(1 + \frac{P}{\sigma^2}) \to 0 \text{ at a rate of } \frac{m}{n} \tag{19.1}$$

Clearly, privacy guarantees improve as m gets smaller, but how small can m be while guaranteeing some utility?

In [Zhou-Lafferty-Wasserman], the authors characterize the utility of doing regression using data privatized via noisy random projections. They consider the compressed linear regression model $Y = AX\beta + \epsilon$ where β is of dimension p and s-sparse, and X is of dimension $n \times p$. They show that if $m = s^2 \log(np)$, then $MSE \to 0$ and $supp(\beta) = supp(\widehat{\beta})$ using lasso regression (under standard incoherence assumptions on design matrix X):

$$\arg\min_{\beta} \frac{1}{n} \|Y - AX\beta\|^2 + \lambda \|\beta\|_1.$$

However, mutual information only preserves privacy on average. A stronger notion of privacy is differential privacy that we discuss next.

19.4.2 Differential Privacy

Differential privacy is a mathematical formalism for a privacy-preserving algorithm. We say an algorithm is (ϵ, δ)-differentially private if for all inputs X, X' differing in at most one value, and for all possible outcomes S:

$$Pr[\mathcal{A}(x) \in S] \leq e^{\epsilon} Pr[\mathcal{A}(x') \in S] + \delta \tag{19.2}$$

where $\mathcal{A}$ refers to the algorithm under consideration.

19.4.2.1 Noisy Random projection approach

One can use noisy random projections to achieve differential privacy. If we let $Y = AX + Z$, where X is the original data matrix and Z has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, then we can achieve (ϵ, δ) differential privacy as long as:

$$\sigma \geq (\max_j \|a_j\|_2) \frac{\sqrt{2(\log \frac{1}{2\delta} + \epsilon)}}{\epsilon} \tag{19.3}$$

where a_j are the columns of the matrix A (see Theorem 1 of [Kenthapadi et al 2012]). The key idea is to show that adding Gaussian noise to the output with $\sigma \geq \Delta_2 \frac{\sqrt{2(\log \frac{1}{2\delta} + \epsilon)}}{\epsilon}$ preserves (ϵ, δ) differential privacy, where Δ_2 is the ℓ_2-sensitivity of the output, i.e. the largest relative change in the output ℓ_2 norm caused by

changing a single entry of the input. This is known as the **Gaussian mechanism**. The result now follows by observing that the largest change in output for two inputs X, X' that differ in one-entry bounded by 1 is $\|A(X - X')\|_2 \leq \max_j \|a_j\|_2$.

References

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

[Zhou-Lafferty-Wasserman] S. Zhou, J. Lafferty and L. Wasserman, "Compressed and Privacy-Sensitive Sparse Regression", IEEE Transactions on Information Theory, vol. 55, no. 2, February 2009.

[Kenthapadi et al 2012] K. Kenthapadi, A. Korolova, I. Mironov and N. Mishra, "Privacy via the Johnson-Lindenstrauss Transform", Journal of Privacy and Confidentiality, 5(1):3971, 2013.