Logistics

Review

Class Road Map - IT-I

- L19 (1/6): Overview, Communications, Gaussian Channel
- L20 (1/8): Gaussian Channel, band limitation, parallel channels, optimization and duality
- L21 (1/13): parallel channels, colored noise, feedback, matrix inequalities
- L22 (1/15): matrix inequalities, rate distortion.
- – (1/20): Monday holiday
- L23 (1/22): rate distortion for Bernoulli, Gaussian, and Multiple Gaussians with unequal noise
- L24 (1/27): main rate distortion theorem, geometry
- L25 (1/29): computing $R(D)$
- L26 (2/3):
- L27 (2/5):
- L28 (2/10):
- L29 (2/12):
- – (2/17): Monday, Holiday
- L30 (2/19):
- L31 (2/24):
- L32 (2/26):
- L33 (3/3):
- L34 (3/5):
- L35 (3/10):
- L36 (3/12):

Cumulative Outstanding Reading

- Read Ch. 10 in our book (Cover & Thomas, “Information Theory”).
- Read Ch. 17 in our book (Cover & Thomas, “Information Theory”) on matrix inequalities.
- Read Ch. 9 in our book (Cover & Thomas, “Information Theory”)
- Read Ch. 5 in Boyd and Vandenberghe’s Convex Optimization book
- Read all readings assigned in EE514a, Fall 2013. (see later lectures on our previous web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/)).

Additional Reading on Rate-Distortion Theory

- “Information Geometry and Alternating Minimization Procedures”, Csiszár & Tusnády, 1983
Homework

- No current outstanding HW.

Announcements

- Office hours on Mondays, 3:30-4:30.
- As always, email me if you want to skype/google hangout rather than come to office hours, also at different times.
On Final Presentations

- Your task is to give a 10-15 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
- The papers must not be ones that we covered in class, although they can be related.
- You need to do the research to find the papers yourself (i.e., that is part of the assignment).
- The majority of the papers must have been published in the last 10 years (so no old or classic papers).
- Your grade will be based on how clear, understandable, and accurate your presentation is (and also milestones).
- This is a real challenge and will require significant work! Many of the papers are complex. To get a good grade, you will need to work very hard to present very complex ideas in an extremely simple yet still precise way.
- Again, don’t expect this to be easy, you might need to try a few topics until you find one that is suitable.

Final Presentation Milestones

All submissions done in PDF file format via our assignment dropbox (https://canvas.uw.edu/courses/880971/assignments)

- Monday, Feb 17th, 11:45pm: Candidate proposed papers submitted. Include short at most 1-page writeup: 1) why you chose these papers; 2) how they are related to each other; 3) why they are important to pure IT; and 4) how they are fundamental and/or deep, and 5) how will you summarize them in a simple and precise way.
- Monday, Feb 24th 11:45pm: Updated list of proposed papers decided, based on feedback. Updated writeup with more description.
- Monday, March 3rd 11:45pm: progress report (at most 1 page). Any background papers you needed to read to better understand your core set. Thoughts on coherent and simple unifying presentation.
- Monday, March 10th, 11:45pm: updated short (≤ 1 page) writeup on more details of how you will present the ideas in a simple fashion.
- Final presentations: Monday, March 17, 2014, 2:30–4:20pm, LOW 102. What to turn in: your slides and a short at most 4 page summary of the papers.
Key Theorem

**Theorem 25.2.1**

Let $R(D)$ be the rate-distortion function and let $R^{(I)}(D)$ be the information rate distortion function. Then

$$R(D) = R^{(I)}(D)$$  \hspace{1cm}(25.18)

- This means that the minimum coding rate for achieving distortion $D$ is, perhaps now unsurprisingly, $R^{(I)}(D)$.
- Two things to prove: (1) that if $(R, D)$ is achievable, then $R > R^{(I)}(D)$, and (2) if $R > R^{(I)}(D)$, then there exists a sequence of codes that can achieve rate-distortion pair $(R, D)$.
- Like in channel capacity and entropy compression case, what happens at $R = R^{(I)}(D)$ depends on the very specific case that one is analyzing.
- Before proving this key theorem, let’s look at Gaussian sources.
Typicality lives

Definition 25.2.2 (distortion $\epsilon$-typical)

Let $p(x, \hat{x})$ be a joint distortion, $d(x, \hat{x})$ a distortion. For any $\epsilon > 0$, $(x^n, \hat{x}^n)$ (a pair of sequences) is distortion $\epsilon$-typical if all four of the below are true:

\begin{align*}
| - \frac{1}{n} \log p(x^n) - H(X) | < \epsilon & \quad \text{$x$-typical} \\
| - \frac{1}{n} \log p(\hat{x}^n) - H(\hat{X}) | < \epsilon & \quad \text{$\hat{x}$-typical} \\
| - \frac{1}{n} \log p(x^n, \hat{x}^n) - H(X, \hat{X}) | < \epsilon & \quad \text{jointly typical} \\
|d(x^n, \hat{x}^n) - Ed(X, \hat{X})| \leq \epsilon & \quad \text{new, “distortion typical”}
\end{align*}

Any $x$ s.t. Equations (25.72)-(25.14) are true define the set $A_{d,\epsilon}^{(n)} \subseteq A_{\epsilon}^{(n)}$.

Probability of typicality

Lemma 25.2.2

\textit{Let $(x_i, \hat{x}_i) \sim p(x, \hat{x})$. Then $\text{Pr}(A_{d,\epsilon}^{(n)}) \to 1$ as $n \to \infty$.}

Proof.

Simple application of the weak law of large numbers, just like before.

Note, this is the same as earlier, except for the distortion but since $d(x^n, \hat{x}^n) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$, we see that $d(x^n, \hat{x}^n) \to Ed(X, \hat{X})$ by the w.l.l.n. as well.
Theorem 25.2.2

\[ \forall (x^n, \hat{x}^n) \in A_d(n), \text{ we have} \]

\[ p(\hat{x}^n) \geq p(\hat{x}^n|x^n)2^{-n(I(X;\hat{X})+3\epsilon)} \]  \hspace{1cm} (25.16)

Proof.

\[ \forall (x^n, \hat{x}^n) \in A_d(n), \text{ we have} \]

\[ p(\hat{x}^n|x^n) = \frac{p(\hat{x}^n, x^n)}{p(x^n)} = \frac{p(\hat{x}^n, x^n)}{p(x^n)p(\hat{x}^n)} \]  \hspace{1cm} (25.17)

\[ \leq \frac{p(\hat{x}^n)}{2^{-n(H(X;\hat{X})+\epsilon)}} \]  \hspace{1cm} (25.18)

\[ = p(\hat{x}^n)2^{n(I(X;\hat{X})+3\epsilon)} \]  \hspace{1cm} (25.19)

Theorem 25.2.2

For \(0 \leq x, y \leq 1\) and \(n > 0\), we have

\[ (1 - xy)^n \leq 1 - x + e^{-yn} \]  \hspace{1cm} (25.16)

Proof.

- \(f(y) \triangleq e^{-y} - 1 + y \Rightarrow f(0) = 0.\)
- and \(f'(y) = -e^{-y} + 1 > 0\) for all \(y > 0.\)
- Thus, \(f(y) > 0\) for all \(y > 0.\)
- \(\Rightarrow\) for \(0 \leq y \leq 1\), we have \(0 \leq 1 - y \leq e^{-y},\) which is a variational lower bound.
**Geometry of Rate Distortion**

- Source $X \sim \mathcal{N}(0, \sigma^2)$.
- A $(2^{nR}, n)$ code with distribution $< D$ is a set of $2^{nR}$ sequences in $\mathcal{R}^n$ s.t. most $x^n$ are “near” a codeword.
- Sources live in sphere of radius $(n\sigma^2)^{1/2}$.
- Goal is to use the fewest number of codewords s.t. every source sequence $X^n$ is within $\sqrt{nD}$ of some codeword.
- Minimum number of such words is

$$\approx \left[ \frac{(n\sigma^2)^{1/2}}{(nD)^{1/2}} \right]^n = \left( \frac{\sigma^2}{D} \right)^{n/2} = 2^{nR(D)} \quad (25.41)$$

for Gaussian sources.

---

**Geometry of Rate Distortion**

- Let $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$ be a large sphere.
- $h(\hat{x}|X) = h(Z) = \frac{1}{2} \log(2\pi e D)$ is a small sphere, and is a region corresponding to codeword $\hat{x}$ in the form of a log volume.
- To make sure that no $x^n$ is too far away from codeword, need to spread out (or cover) the large volume as much as possible.
- The minimum number of such codewords is then:

$$\geq \frac{2^n h(X)}{2^n h(Z)} = \frac{2^n \log(2\pi e \sigma^2)}{2^n/2 \log(2\pi e D)} = \left( \frac{\sigma^2}{D} \right)^{n/2} = 2^{nR(D)} \quad (25.41)$$

- Again, meaning $R = \frac{1}{2} \log(\sigma^2 / D)$ bits per source symbol to compress with distortion $D$. 

General Multivariate Normal

- Suppose $X \sim \mathcal{N}(0, C)$ where $C$ is general positive definite matrix.
- Then $R(D)$ can be easily and analytically calculated by using the reverse water filling argument on the eigenvalues of $C$.
- This is analogous to the water filling argument for computing the channel capacity of a general Gaussian.

Computing $R(D)$

- We’ve seen that for certain special cases (e.g., Bernoulli sources, Gaussian sources, and now also for Gaussian vector w. full covariance matrix), we can compute $R(D) = R^{(I)}(D)$ exactly.
- What if we do not have such simple properties of $X$?
- Let $X \sim p(x)$ which is over multi-alphabet $\mathcal{X}$.
- Goal: compute in general

$$R(D) = \min_{q(\hat{x}|x) \sum_{x, \hat{x}} p(x)q(\hat{x}|x)d(x, \hat{x}) \leq D} I(X; \hat{X}) \quad (25.1)$$
Computing $R(D)$

- Can restate this problem as:

$$R(D) = \min_{q(\hat{x}|x)} I(X; \hat{X}) \quad (25.2)$$

s.t.  
$$q(\hat{x}|x) \geq 0 \ \forall \hat{x}, x \quad (25.3)$$

$$\sum_{\hat{x}} q(\hat{x}|x) = 1 \ \forall x \quad (25.4)$$

$$\sum_{\hat{x}, x} q(\hat{x}|x)p(x)d(x, \hat{x}) = D \quad (25.5)$$

where

$$I(X; \hat{X}) = \sum_{x, \hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{q(\hat{x})} \quad (25.6)$$

and

$$q(\hat{x}) = \sum_{x} p(x)q(\hat{x}|x) \quad (25.7)$$

Marginalization vs. Projection

- We’re going to see that the marginalization $q(\hat{x}) = \sum_{x} p(x)q(\hat{x}|x)$ can be viewed as a form of projection.
- Generic projection. We have distance measure $d(\cdot, \cdot)$ and a constraint set $\mathcal{P}$, and a vector $\hat{x} \notin \mathcal{P}$.
- We want to find the member of $\mathcal{P}$ that is closest to $\hat{x}$ where closeness is measured via $d(\cdot, \cdot)$, i.e.,

$$x^* \in \arg\min_{x \in \mathcal{P}} d(\hat{x}, x) \quad (25.8)$$
Computing $R(D)$

- Both inequality (Eq.(25.3)) and equality constraints (Eqs.(25.4) & (25.5)).
- We have convex objective in $q(\hat{x}|x)$ for fixed $p(x)$.
- Q: Why ok to equal $D$ and not $\leq D$ in the above? A: we know can only make $R(D)$ smaller by making $D$ larger.
- For the moment, let’s ignore the inequality constraint $q(\hat{x}|x) \geq 0$ and hope that we find everywhere positive solutions.
- We get objective (Lagrangian) in the form:

$$J(q) = \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda(\sum_{x,\hat{x}} p(x)q(\hat{x}|x)d(x, \hat{x}) - D) + \sum_\nu(x) \left( \sum_{\hat{x}} q(\hat{x}|x) - 1 \right)$$

(25.9)

Eq (25.9) satisfies Slater’s conditions, so strong duality holds, and the objective and all constraints are differentiable.

Thus, happily, the KKT conditions hold, so we need

$$\frac{\partial J}{\partial q(\hat{x}|x)} = 0 \Rightarrow |\mathcal{X}| \times |\hat{\mathcal{X}}| \text{ equations} \quad (25.10)$$

The equality constraints give us an additional $|\mathcal{X}| + 1$ equations.

How many unknowns? $|\mathcal{X}| \times |\hat{\mathcal{X}}|$, so this should be solvable (if feasible, and assume is).

If solution is such that $q(\hat{x}|x) > 0$ for all $x, \hat{x}$ then we have solved the problem since this solution is a minimum without requiring the inequality constraints (and we still have equality constraint $\sum_{\hat{x}} q(\hat{x}|x) = 1$ so solution is feasible).
When might this occur? Small $D$ requires a high rate, meaning we need to use all of the symbols in $\hat{X}$.
- It might even be that for all $0 \leq D \leq D_{\text{max}}$, we use all symbols.
- If, on the other hand, there are one or more $q(\hat{x}|x)$ that are negative, then the min of $I(X; \hat{X})$ falls outside of the probability simplex (defined as $\triangle q(\hat{x},x)$).
- $\triangle q(\hat{x},x)$ is convex, so in this case, the minimization with the constraint must lie on a boundary of this convex set (essentially KKT conditions).

General result: convex minimization over a convex set, an update to minimize objective without constraint need only be projected back to nearest point in convex set.
- So, if solution of problem without inequality constraint is negative (say $q(\hat{x}'|x) < 0$ for some $\hat{x}'$), then projection back onto convex $\triangle q(\hat{x},x)$ at boundary will set it to zero.
- That is, we get that $q(\hat{x}'|x) = 0$
- This means we can, for this $x$, essentially “remove” alphabet symbol $\hat{x}'$ (or perhaps several such symbols) and form a compressed alphabet $\hat{X}'$ where $|\hat{X}'| < |\hat{X}|$
- This happens when $D$ is big, and by minimizing the rate, we don’t need to “use up” all the symbols in $\hat{X}$.
- Thus, lets first solve the partially unconstrained problem (i.e., the one without the inequality constraints, Eq.(25.3)).
Computing $R(D)$

- First, define $\mu(x) \triangleq e^{-\nu(x)/p(x)}$ so that $-\log \mu(x) = \nu(x)/p(x)$.
- The Lagrangian objective (Eq (25.9)) can be stated, ignoring constants that cancel out when taking derivatives w.r.t. $q(\hat{x}|x)$, as:

\[
J(q) = \sum_{x, \hat{x}} p(x)q(\hat{x}|x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})} + \lambda d(x, \hat{x}) + \frac{\nu(x)}{p(x)} \right]
\]

(25.11)

\[
= \sum_{x, \hat{x}} p(x)q(\hat{x}|x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right]
\]

(25.12)

\[
= \sum_{x, \hat{x}} p(x)q(\hat{x}|x)\beta(x, \hat{x})
\]

(25.13)

where

\[
\beta(x, \hat{x}) = \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right]
\]

(25.14)

After some algebra, we get:

\[
\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] = 0
\]

(25.15)

Implying that

\[
q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x, \hat{y})}}
\]

(25.16)

This expresses $q(\hat{x}|x)$ in terms of $q(\hat{x})$, so if we can solve for $q(\hat{x})$, then we can get $q(\hat{x}|x)$. We first do a little intuition.
Computing $R(D)$

- The update is:

$$q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \tag{25.17}$$

- Note that to make this a valid normalized distribution, we must take $\mu(x) = \sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}$ since $\sum_{\hat{x}} q(\hat{x}|x) = 1$.

- If $d(x, \hat{x})$ is large, then $q(\hat{x}|x)$ will be small. Makes sense that we don’t in general want to use $\hat{x}$ for $x$ if distortion is large.

- This, however, is balanced by overall $q(\hat{x})$ which will force us to start using $\hat{x}$ for $x$ if $q(\hat{x})$ is large.

To solve for $q(\hat{x}) > 0$, we find $q(\hat{x}) = \sum_{x} p(x)q(\hat{x}|x)$, yielding:

$$\tilde{q}(\hat{x}) = \sum_{x} p(x) \left( \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \right) \tag{25.18}$$

$$= \tilde{q}(\hat{x}) \frac{\sum_{x} p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \tag{25.19}$$

- So, for all $\hat{x}$ such that $q(\hat{x}) > 0$ we have

$$C(\hat{x}) = \sum_{x} \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} = 1 \tag{25.20}$$

- Thus, if $q(\hat{x}) > 0$ for all $\hat{x}$, then this defines $|\hat{x}|$ simultaneous equations $\{C(\hat{x}) = 1\}_{\forall \hat{x}}$ which, along with the distortion constraint equation, can be used to solve the $|\hat{x}|$ unknown quantities $\{q(\hat{x})\}_{\forall \hat{x}}$, for the current $\lambda$. 

According to Eq. (25.17), as long as \( q(\hat{x}) > 0 \) then \( q(\hat{x}|x) > 0 \) as well.

So, we can then choose a \( \lambda \) and use it to compute particular point on the \( R(D) \) curve.

For a given set of values \((\lambda, \{q(\hat{x})\})\), we have

\[
D = \sum_{x, \hat{x}} \frac{p(x)}{\mu(x)} q(\hat{x}) e^{-\lambda d(x, \hat{x})} = D(\lambda, \{q(\hat{x})\}) \tag{25.21}
\]

It can also be shown, moreover, that

\[
R = -(\lambda D + \sum_x p(x) \log \mu(x)) = R(\lambda, \{q(\hat{x})\}) \tag{25.22}
\]

\[
= sD + \sum_x p(x) \log 1/\mu(x) \quad \text{where } s = -\lambda \tag{25.23}
\]

Since \( s = -\lambda \) determines \( D \), if \( s \) yields a large enough \( D \) we will ultimately get some cases where \( q(\hat{x}) \leq 0 \).

In fact, this is sufficient to eliminate all instances of \( \hat{x} \) as we now further show.
Computing $R(D)$

Theorem 25.4.1

$\forall s > -\infty$, for optimal $q(\hat{x})$, if $q(\hat{x}|x) = 0$ for any one $x$ then $q(\hat{x}|x) = 0$ for all $x$. Thus, that particular $\hat{x}$ may be deleted from the alphabet.

Proof.

- Lets bring the inequality constraints back in for $q(\hat{x}|x) \geq 0$ for a particular $(\hat{x}, x)$ pair:

$$L(q) = J(q) + \gamma q(\hat{x}|x) \quad (25.24)$$

- Then setting $\frac{\partial L}{\partial p(\hat{y}|x)} = 0$, we get the following relation

$$q(\hat{y}|x) = \begin{cases} 
q(\hat{y}) e^{sd(x,\hat{y})} & \text{if } \hat{y} \neq x \\
q(\hat{y}) e^{(sd(x,\hat{y})+\gamma/p(x))} & \text{if } \hat{y} = x 
\end{cases} \quad (25.25)$$

Note that the above inequality is strict since $e^{f(\hat{y},x)} > 0$ and $q(\hat{y}) \geq 0$.

- Thus, $\mu(x) = \sum \hat{y} q(\hat{y}) e^{f(\hat{y},x)} > 0$

$$\mu(x) = \sum \hat{y} q(\hat{y}) e^{f(\hat{y},x)} > 0 \quad (25.26)$$

- Note that the above inequality is strict since $e^{f(\hat{y},x)} > 0$ and $q(\hat{y}) \geq 0$.

- If $q(\hat{x}|x) = 0$, it must be due to $q(\hat{x}) = 0$ and not due to either $\mu(x)$ or to the $e^{f(\hat{y},x)}$ factor, as long as $s > -\infty$.

- But if $q(\hat{x}) = 0$ then we have $q(\hat{x}|x) = 0$ for all $x$. 

Computing $R(D)$

- More intuition: From previous definition, we have
  \[ q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum \hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \]  
  \[ (25.27) \]

- If $q(\hat{x}|x) = 0$ for some $\hat{x}$, then this must be due to $q(\hat{x}) = 0$ since nothing else in the definition can be 0.

- We also have a nice meaning for $s = -\lambda$.

Theorem 25.4.2

The parameter $s = -\lambda$ represents the slope of the rate-distortion function at the point $(D_s, R_s)$ that one generates parametrically from the parametric form above. I.e.

\[ R' = \left. \frac{dR}{dD} \right|_{D_s} = s \]  

(25.28)

Proof.

Take derivatives and use the chain rule . . .

- Pictorially,
Thus, we have a way to compute \( R(D) \) in principle for any \( s = -\lambda \).

To get the resulting distribution, we need to find the \( q(\hat{x}) \) values, and if \(< 0\) remove symbols, and repeat.

We continue this process until all are positive.

If we have only one left, then we have a \( R = 0 \) case.

Also, solution to the set of equations might be hard (or an analytical solution might not exist).

Fortunately, there is a better way to do all of this, as we now proceed to show.

---

Consider the problem: we have two convex sets \( A, B \subseteq \mathbb{R}^n \).

We have a distance (e.g., Euclidean, or 2-norm) \( d(a, b) \).

Goal is to form:

\[
\begin{align*}
    d_{\text{min}} &= \min_{a \in A, b \in B} d(a, b) \quad (25.29)
\end{align*}
\]

Consider the following algorithm:

1. Choose \( a_0 \in A \) arbitrarily;
2. for \( n = 1 \ldots \) do
3. Choose \( b_n \in \arg\min_{b \in B} d(a_{n-1}, b) \);
4. Choose \( a_n \in \arg\min_{a \in A} d(a, b_n) \);
Alternating Minimization Between 2 convex sets

- with notation switch $A, B$ correspond to sets $\mathcal{P}, \mathcal{Q} \subseteq \mathbb{R}^n$.
- Jump to animation slides. [Click Here]

In the next lecture, we will prove the conditions under which it will be the case that

$$\lim_{n \to \infty} d(a_n, b_n) = d(a^*, b^*)$$ (25.30)

where

$$(a^*, b^*) = \arg\min_{a,b} d(a, b)$$ (25.31)

and

$$\lim_{n \to \infty} a_n = a^*, \quad \lim_{n \to \infty} b_n = b^*$$ (25.32)

For now, let's assume that it works for both rate distortion and channel capacity, and derive these cases.
Theorem 25.5.1

Let \( p(x, y) = p(x)p(y|x) \). Then

1. If \( r^*(y) = \sum_x p(x)p(y|x) \), then

   \[
   D(p(x)p(y|x)||p(x)r^*(y)) = \min_{r(y) \in \Delta} D(p(x)p(y|x)||p(x)r(y))
   \]

   (25.33)

2. If \( r^*(x|y) = \frac{p(x)p(y|x)}{\sum_x p(x)p(y|x)} = p(x|y) \) then

   \[
   \max_{r(x|y) \in \Delta^2} \sum_{x,y} p(x)p(y|x) \log \frac{r(x|y)}{p(x)} = \sum_{x,y} p(x)p(y|x) \log \frac{r^*(x|y)}{p(x)}
   \]

   (25.34)

Another way of seeing these theorems is that we have:

\[
I(X; Y) = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}
\]

(25.35)

\[
= \min_{r(y)} \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)r(y)}
\]

(25.36)

\[
= \max_{r(x|y)} \sum_{x,y} p(x, y) \log \frac{r(x|y)p(y)}{p(x)p(y)}
\]

(25.37)
Back to rate distortion

Therefore, we can write the rate-distortion function as

\[
R(D) = \min_{q(\hat{x}|x): \sum_{x,\hat{x}} p(x)q(\hat{x}|x)d(x,\hat{x}) \leq D} I(X; \hat{X})
\]

\[
= \min_{q(\hat{x}|x): \sum_{x,\hat{x}} p(x)q(\hat{x}|x)d(x,\hat{x}) \leq D} \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{q(\hat{x})}
\]

\[
= \min_{q(\hat{x}|x): \sum_{x,\hat{x}} p(x)q(\hat{x}|x)d(x,\hat{x}) \leq D} \min_{r(\hat{x})} \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{r(\hat{x})}
\]

\[
= \min_{r(\hat{x})} \min_{q(\hat{x}|x): \sum_{x,\hat{x}} p(x)q(\hat{x}|x)d(x,\hat{x}) \leq D} \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{r(\hat{x})}
\]

\[
= \min_{r(\hat{x})} \left( \log \left( \frac{q(\hat{x}|x)}{r(\hat{x})} \right) \right)
\]

\[
= \min_{r(\hat{x})} \left( \log \left( \frac{q(\hat{x}|x)}{r(\hat{x})} \right) \right)
\]

\[
= -I(X; \hat{X})
\]
The last switch of minimizations can be done since:
1. $-\log r(\hat{x})$ is convex in $r(\hat{x})$ for fixed other parameters
2. $x \log x$ is convex in $x$
3. minimization over convex sets of the form $Ap \leq D$.
All of which allows us to swap the mins.
This then gives $R(D)$ in the alternating minimization form:

$$R(D) = \min_{q \in B} \min_{p \in A} D(p||q)$$

(25.46)

where

$$A = \{q(x, \hat{x}) : q(x, \hat{x}) = p(x)r(\hat{x}) \text{ for arbitrary } r(\hat{x})\}$$

(25.47)

$$B = \left\{ p(x, \hat{x}) : p(x, \hat{x}) = q(\hat{x}|x)p(x) \text{ s.t. } \sum_{x,y} p(x, \hat{x})d(x, \hat{x}) \leq D \right\}$$

(25.48)

So, to compute $R(D)$ at some point $s = -\lambda$, start with some arbitrary $r(\hat{x})$, and find the corresponding $q(\hat{x}|x)$.

From earlier, we have that

$$q(\hat{x}|x) = \frac{r(\hat{x})e^{-\lambda d(x, \hat{x})}}{\sum_{\hat{y}} r(\hat{y})e^{-\lambda d(x, \hat{y})}}$$

(25.49)

Once we have $q(\hat{x}) = q(\hat{x}|x)p(x)$, we find corresponding next $r(\hat{x})$ from the projection

$$r(\hat{x}) = \sum_{x} p(x)q(\hat{x}|x)$$

(25.50)

We repeat this alternating projection/minimization procedure until convergence.
This will converge to $R(D)$ at $s$. 

Computing $R(D)$
Recall channel capacity, where we have a noisy channel, a capacity $C$, and Shannon’s theorem saying we can only communicate with vanishingly small probability of error if $R < C$.

In this case, we have:

$$C = \max_{Q} \max_{P} \sum_{x,y} r(x)p(y|x) \log \frac{q(x|y)}{r(y)}$$  \hspace{1cm} (25.51)

We guess, starting $r(x)$ and then iterate the following two equations:

$$q(x|y) = \frac{r(x)p(y|x)}{\sum_{x} r(x)p(y|x)}, \quad r(x) = \frac{\prod_{y} q(x|y)p(y|x)}{\sum_{x} \prod_{y} q(x|y)p(y|x)}$$  \hspace{1cm} (25.52)

**Summary**

Let $\mathcal{P}, \mathcal{Q}$ be convex sets of finite measures, meaning for each $P \in \mathcal{P}$, $\sum_{x} p(x) = 1$, and for all $x \in \mathcal{X}$, $p(x) \geq 0$.

- Define $P_{n} \in \mathcal{P}$ arbitrarily.
- Define $Q_{n} \in \text{argmin}_{Q \in \mathcal{Q}} D(P_{n}||Q)$.
- That is, we have the following procedure:

$$Q_{n} \in \text{argmin}_{Q \in \mathcal{Q}} D(P_{n}||Q)$$  \hspace{1cm} (25.53)

$$P_{n+1} \in \text{argmin}_{P \in \mathcal{P}} D(P||Q_{n})$$  \hspace{1cm} (25.54)

Then the result we will get is that:

$$D(P_{n}||Q_{n}) \to \inf_{(P,Q) \in (\mathcal{P}_{0},\mathcal{Q})} D(P||Q)$$  \hspace{1cm} (25.55)

where $\mathcal{P}_{0} = \{P \in \mathcal{P} : D(P||Q_{n}) < \infty \text{ for some } n\}$ and $P_{n} \to P^{*}$, $Q_{n} \to Q^{*}$ sometimes as well.

$\mathcal{P}_{0}$ are the entries of $\mathcal{P}$ that we care about.
**Summary**

- This process has a geometric flavor, since it corresponds to alternating “projections” based on treating KL as a “distance” in some odd sense.
- It also generalizes (and offers guarantees for) a number of problems, including:
  - Maximum likelihood estimation for mixtures, hidden Markov models, and other graphical models (i.e. the expectation-maximization or EM algorithm).
  - Computing rate-distortion function (Blahut-Arimoto algorithm).
  - Computing the channel capacity function.
  - Optimal investment portfolios.
  - Many semi-supervised learning objectives in machine learning (including forms of “label propagation”, “measure propagation”, etc.).
- The application depends on the quasi-distance \(d(P, Q)\) where \(d : \mathcal{P} \times \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}\) which need not be KL-divergence.

**Properties of \(d\)**

- Let \(d(P, Q)\) be a half extended-real valued function. That is, for \(P \in \mathcal{P}, Q \in \mathcal{Q}\), we have \(d(P, Q) > -\infty\) (we exclude \(-\infty\) but allow \(\infty\)).
- Also, \(d(P, Q') = \min_{Q \in \mathcal{Q}} d(P, Q) < \infty\). This minimization is denoted as \(P \overset{1}{\to} Q'\) where we are holding \(P\) fixed (“1” indicates that \(P\), the first argument of \(d\), is being held fixed) and minimizing the second argument down to \(Q'\).
- Similarly, \(d(P', Q) = \min_{P \in \mathcal{P}} d(P, Q) < \infty\) is denoted \(Q \overset{2}{\to} P'\), indicating we minimize over \(P\), holding the 2nd argument \(Q\) fixed.
- Sequences obtained by alternating minimization \(\{(P_n, Q_n)\}_{n=0}^{\infty}\) as:
  \[
P_0 \overset{1}{\to} Q_0 \overset{2}{\to} P_1 \overset{1}{\to} Q_1 \overset{2}{\to} P_2 \overset{1}{\to} Q_2 \overset{2}{\to} P_3 \overset{1}{\to} Q_3 \overset{2}{\to} \ldots
  \]
  where we start arbitrarily with \(P_0\).
- Goal: sufficient conditions for the convergence of the alternating minimization procedure.
Five Points Property

Definition 25.6.1 (Five Points Property (5PP))

For a \( P \in \mathcal{P} \), the quasi-distance \( d : \mathcal{P} \times \mathcal{Q} \to \mathbb{R} \cup \{+\infty\} \) satisfies the five points property if: \( \forall Q \in \mathcal{Q}, \forall Q_0 \in \mathcal{Q} \), we have:

\[
d(P, Q) + d(P, Q_0) \geq d(P, Q_1) + d(P_1, Q_1)
\]

whenever \( Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \). \( d(\cdot, \cdot) \) satisfies 5PP if it satisfies 5PP for all \( P \in \mathcal{P} \).

- Note: this is a property of a quasi-distance (or divergence) across sets \( \mathcal{P} \) and \( \mathcal{Q} \).
- It is a definition on sets of 5 points! (obviously \( \odot \)).
- Compare triangle inequality: We have one set, say, \( \mathcal{P} \). Triangle inequality would require that for all triples of points \( P_1, P_2, P_3 \in \mathcal{P} \),

\[
d(P_1, P_2) + d(P_2, P_3) \geq d(P_1, P_3),
\]

where in this case \( d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+ \).
Properties

- We will prove that if five points property holds (either $\forall P \in \mathcal{P}$, or some other conditions that are specified later), then

$$\lim_{n \to \infty} d(P_n, Q_n) = \inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} d(P, Q) = d_{\min}$$  \hspace{1cm} (25.58)

as long as

$$d_{\min} = \inf_{P \in \mathcal{P}_0, Q \in \mathcal{Q}} d(P, Q)$$  \hspace{1cm} (25.59)

where

$$\mathcal{P}_0 = \{P : P \in \mathcal{P}, d(P, Q_n) < \infty \text{ for some } n\}$$  \hspace{1cm} (25.60)

- Note, $\mathcal{P}_0$ depends on the sequence, of course, and $\mathcal{P}_0 = \mathcal{P}$ if $d$ is finite valued.

Definitions

- We define, for $A \subseteq \mathcal{P}$ and $B \subseteq \mathcal{Q}$,

$$d(A, B) \triangleq \inf_{P \in A, Q \in B} d(P, Q)$$  \hspace{1cm} (25.61)

Since $d(P, Q) \in \mathbb{R} \cup \{+\infty\}$, $d(A, B)$ does not take the value $-\infty$.

Lemma 25.6.2

Let $\{(P_n, Q_n)\}_{n=0}^\infty$ be sequences (not necessarily generated via alternating minimization). Then

$$d(P_n, Q_n) \geq d(\mathcal{P}_0, Q) \quad \forall n$$  \hspace{1cm} (25.62)

Proof.

Obvious via definitions.

Our goal is to first find when $\lim_{n \to \infty} d(P_n, Q_n) = d(\mathcal{P}_0, Q)$. 
recall \( \limsup / \liminf \)

- Recall,
  \[
  \limsup_{n \to \infty} a_n \overset{\Delta}{=} \inf_{n>0} \left( \sup_{k>n} a_k \right) = \inf S \tag{25.63}
  \]
  where
  \[ S = \{ a : a = \sup B_n \text{ for some } n, \text{ with } B_n = \{a_n, a_{n+1}, \ldots \} \}. \]

- For example, while \( \lim_{x \to \infty} \sin(x) \) does not exist,
  \( \limsup_{x \to \infty} \sin(x) = 1. \)

- Also, \( \limsup_{x \to \infty} (\sin(x) - \sin^2(x)) = 1/4. \)

- Thus, \( \limsup \) allows for oscillation in the sequences and in some sense \( \limsup \) asks for infimum convergence in the local maxima (or perhaps better, “reverse-time cumulative” local maxima).

- Also,
  \[
  \liminf_{n \to \infty} a_n \overset{\Delta}{=} \sup_{n>0} \left( \inf_{k>n} a_k \right) \tag{25.64}
  \]
  so \( \liminf \) asks for supremum convergence in the local minima.

**Key Lemma**

**Lemma 25.6.3**

Let \( a_n, b_n \) for \( n = 0, 1, \ldots \) be extended real sequences in the sense \( \forall n, a_n, b_n \in \mathbb{R} \cup \{+\infty\}. \) Let \( c \) be finite arbitrary such that:
\[
c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots. \tag{25.65}
\]
And also assume that
\[
\limsup_{n \to \infty} b_n > -\infty, \quad \text{and} \quad \exists n_0 \text{ s.t. } b_{n_0} < \infty. \tag{25.66}
\]
Then
\[
\liminf_{n \to \infty} a_n \leq c \tag{25.67}
\]
Also, if in addition, we assume that
\[
\sum_{n=0}^{\infty} (c - a_n)^+ < \infty \quad \text{then} \quad \sum_{n=n_0+1}^{\infty} |a_n - c| < \infty \tag{25.68}
\]
and as a result
\[
\lim_{n \to \infty} a_n = c \tag{25.69}
\]
**Key Lemma**

**Proof.**

- First, assume case where \( \sum_{n=0}^{\infty} (c - a_n)^+ = \infty \),
- then since \( c \) is finite, for any \( n \) where \( a_n = +\infty \), those \( n \)s don’t contribute since \( (c - \infty)^+ = 0 \). So we may assume \( a_n < \infty \).
- In such case, we are summing finite values and getting an infinite result so \( a_n \) can’t converge to anything strictly greater than \( c \) (i.e., we can’t have that \( \lim_{n \to \infty} a_n > c \) since if so, eventually we’d get \( (c - a_n)^+ \) and the sum would be finite).
- Thus, \( \lim \inf_{n \to \infty} a_n \leq c \).

...
Key Lemma

Proof.

- Then, if \( \sum_{n=n_0+1}^{\infty} (c-a_n)^+ < \infty \) and since in such case \( \sum_{n=n_0+1}^{\infty} (c-a_n) < \infty \), this means that \( \sum_{n=n_0+1}^{\infty} |a_n - c| < \infty \).
- Why? Let \( a^+ = \max(a,0) \) and \( a^- = \max(-a,0) \) so that \( a = a^+ - a^- \) and \( |a| = a^+ + a^- \). All are \( \neq -\infty \). Then if \( a = a^+ - a^- = c_\pm < \infty \) and if \( a^+ = c_+ < \infty \), then \( |a| = a^+ + a^- = c_\pm < \infty \).
- Then when \( \sum_{n=n_0+1}^{\infty} |a_n - c| < \infty \), this means that \( \lim_{n \to \infty} a_n = c \).

...
1st Main theorem

**Theorem 25.6.4**

Given a set of arbitrary sequences \( \{P_n\}_{n=0}^{\infty}, \{Q_n\}_{n=0}^{\infty} \) from (resp.) \( P \) and \( Q \) such that the five-points property holds as follows:

\[
d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n) \quad n = 1, 2, \ldots \tag{25.72}
\]

*Note: no minimization done here, only 5PP condition on the sequences.* Then if either: A) \( \forall P \in P_0 \); or B) for some \( P \in P_0 \) s.t. \( d(P, Q) = d(P_0, Q) \), we have:

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q). \tag{25.73}
\]

And if A holds then \( d(P_n, Q_n) \) is non-increasing. And if B holds then

\[
\sum_{n=1}^{\infty} (d(P_n, Q_n) - d(P_0, Q)) < \infty \tag{25.74}
\]

**Proof of 1st main theorem**

**Proof.**

- If \( P_0 = \emptyset \) then, for all \( n \geq 1 \), we have
  \[
d(P_n, Q_n) = d(P_0, Q) = \inf_{P \in P_0, Q \in Q} d(P, Q) = \inf \emptyset = \infty \tag{25.75}
\]
  so theorem is true in this case (l.h.s. holds by definition of \( P_0 \)).
- Suppose \( P_0 \neq \emptyset \) and that Eq (25.72) holds for some \( P \in P_0 \).
- Then, lemma 25.6.3 \( (c + b_{n-1} \geq b_n + a_n, \ n = 1, 2, \ldots) \) with
  \[
c = d(P, Q), \quad b_n = d(P, Q_n), \quad a_n = d(P_n, Q_n) \tag{25.76}
\]
- Why? Since \( P \in P_0 \), we have both \( \exists n_0 \) s.t., \( b_{n_0} < \infty \) and also \( c < \infty \) (all by the def of \( P_0 \)) implying \( a_n < \infty \) for \( n \geq n_0 \).
- Also, \( \limsup_{n \to \infty} b_n > -\infty \) since Eq (25.72) with \( n = n_0 + 1 \) implies \( c > -\infty \), and \( b_n \geq c = d(P, Q) > -\infty \), since
  \[
d \in \mathbb{R} \cup \{+\infty\}.\]
Proof of 1st main theorem

Proof.

- So, lemma 25.6.3 holds here and from it we get:
  - Under A: \( \forall P \in P_0 \), we have
    \[
    \liminf_{n \to \infty} a_n = \liminf_{n \to \infty} d(P_n, Q_n) \leq c = d(P, Q) < \infty \quad (25.77)
    \]
  - Thus, \( \forall P \in P_0 \), we have that \( d(P_n, Q_n) \) “converges” to a finite value (not \( \infty \)) since \( d > -\infty \) (since it holds for all \( P \in P_0 \), we have \( \liminf_{n \to \infty} d(P_n, Q_n) \leq d(P_0, Q) < \infty \)).
  - Recall, \( d(P_n, Q_n) \geq d(P_0, Q) \) for \( n = 0, 1, \ldots \) for any sequence \( \{P_n, Q_n\}_{n=0}^\infty \)
  - Also, let \( P = P_{n-1} \) in Eq (25.72), so we get:
    \[
    d(P_{n-1}, Q) + d(P_{n-1}, Q_{n-1}) \geq d(P_{n-1}, Q_n) + d(P_n, Q_n) \quad (25.78)
    \]
    \[
    \geq 0
    \]
    \[
    (25.79)
    \]

- This implies that
  \[
  d(P_{n-1}, Q_{n-1}) \geq d(P_{n-1}, Q_n) - d(P_{n-1}, Q) + d(P_n, Q_n) \geq d(P_n, Q_n)
  \]

- So, a non-increasing sequence with a lower bound (even so if \( d(P_{n-1}, Q_{n-1}) = \infty \) or if it is finite) will converge.
- Non-increasing sequence with a lower bound of \( d(P_0, Q) \) means that
  \[
  \lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \quad (25.80)
  \]
Proof of 1st main theorem

Proof.

- Next, under B (for some \( P \in \mathcal{P}_0 \) s.t. \( d(P, Q) = d(\mathcal{P}_0, Q) \)), we have that

\[
d(P_n, Q_n) \geq d(\mathcal{P}_0, Q) = d(P, Q)
\]

which follows since, as mentioned earlier, \( d(P_n, Q_n) \geq d(\mathcal{P}_0, Q) \) for \( n = 0, 1, \ldots \) for any sequence \( \{P_n, Q_n\}_{n=0}^{\infty} \)

- This means,

\[
a_n = d(P_n, Q_n) \geq c = d(P, Q)
\]

or that \( c - a_n \leq 0 \), implying that \( \sum_{n=0}^{\infty} (c - a_n)^+ < \infty \)

- From lemma 25.6.3, this gives \( \lim_{n \to \infty} a_n = c \) or

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P, Q) = d(\mathcal{P}_0, Q)
\]

Proof of 1st main theorem

Proof.

- And since (as shown earlier)

\[
\sum_{n=n_1}^{\infty} (a_n - c) < \infty
\]

we have

\[
\sum_{n=n_1}^{\infty} (d(P_n, Q_n) - d(\mathcal{P}_0, Q)) < \infty
\]

(25.84)

(25.85)

(25.86)

(25.87)
Sequences

- Consider next sequences \( \{(P_n, Q_n)\}_{n=0}^{\infty} \) constructed by alternating minimization with arbitrary starting point \( P_0 \in \mathcal{P} \)

\[
P_0 \Rightarrow Q_0 \Rightarrow P_1 \Rightarrow Q_1 \Rightarrow P_2 \Rightarrow Q_2 \Rightarrow P_3 \Rightarrow Q_3 \Rightarrow \cdots \quad (25.88)
\]

- Then we have that:

\[
d(P_n, Q_n) \geq \min(P_{n+1}, Q_n) \geq \min(P_{n+1}, Q_{n+1}) \quad \text{for } n = 0, 1, \ldots
\]

\[
\text{(25.89)}
\]

- And thus we have an ever non-increasing sequence.

- If 5PP holds for some \( P \in \mathcal{P} \) (for now, do some \( P \) but will later relate it to \( P_0 \)), and if we construct an alternating minimization sequence starting at some \( P_0 \in \mathcal{P} \), we have conditions of Theorem 25.6.4 met at \( P \in \mathcal{P}_0 \)

\[
\text{That is, for } n = 1 \text{ we have } Q_0 \Rightarrow P_1 \Rightarrow Q_1 \text{ so}
\]

\[
d(P, Q) + d(P, Q_0) \geq d(P, Q_1) + d(P_1, Q_1) \quad \forall Q, Q_0
\]

\[
\text{(25.90)}
\]

which is just the 5PP which is presumed to hold.

- Thus, this also certainly holds for \( Q_0 \) such that \( P_0 \Rightarrow Q_0 \).

- and also have the same when the first term is is the particular \( Q \) that achieves \( d(P, Q) = \inf_{Q \in \mathcal{Q}} d(P, Q) \).

- For \( n = 2 \) we have \( Q_1 \Rightarrow P_2 \Rightarrow Q_2 \) so

\[
d(P, Q) + d(P, Q_1) \geq d(P, Q_2) + d(P_2, Q_2) \quad \forall Q, Q_0
\]

\[
\text{(25.91)}
\]

so also true for \( Q_1 \) such that \( P_1 \Rightarrow Q_1 \).

- Same for \( n > 2 \), etc.
So Theorem 25.6.4 holds in this case (i.e., \( \lim_{n \to \infty} d(P_n, Q_n) = d(P, Q) \)).

On the other hand, we want other (perhaps easier) conditions that, if true, imply the five points property.

This will making checking 5PP much easier.

We identify two that, if both hold, will imply 5PP.

These are the three-points property (3PP) and the four-points property (4PP), and 3PP + 4PP = 5PP.

**Three Points Property**

**Definition 25.6.5 (Three Points Property (3PP))**

Let \( \delta(P, P') \geq 0 \) be a function \( \delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+ \) such that \( \delta(P, P) = 0 \) for all \( P \in \mathcal{P} \). For \( d : \mathcal{P} \times Q \to \mathbb{R} \cup \{+\infty\} \) and \( \delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+ \), the three points property for \( P \in \mathcal{P} \) holds if \( \forall Q_0 \)

\[
\delta(P, P_1) + d(P_1, Q_0) \leq d(P, Q_0) \quad \text{whenever} \quad Q_0 \xrightarrow{\cdot} P_1
\]  

(25.92)

So sort of like a reverse triangle inequality.
**Three Points Property**

![Three Points Property Diagram]

- **Definition:**
  \[
  \forall Q_0 \in Q \quad \exists P_1 \in \text{argmin} \quad \delta(P, P_1) + d(P_1, Q_0) \geq d(P, Q_0)
  \]

**Four Points Property (4PP)**

**Definition 25.6.6 (Four Points Property (4PP))**

The 4PP holds for \( P \in \mathcal{P} \) if \( \forall Q \in \mathcal{Q} \), and \( \forall P_1 \in \mathcal{P} \), we have that

\[
d(P, Q_1) \leq \delta(P, P_1) + d(P, Q) \quad \text{whenever} \quad P_1 \xrightarrow{1} Q_1
\]

(25.93)
### Four Points Property (4PP)

**Theorem 25.6.7**

Let \( \{(P_n, Q_n)\}_{n=0}^{\infty} \) be sequences obtained by alternating minimization. Then

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q)
\]  

(25.94)

if \( P \) is defined by either: A) all \( P \in \mathcal{P}_0 \); or B) some \( P \in \mathcal{P}_0 \) with \( d(P, Q) = d(P_0, Q) \) has the 5PP. Also,

1. \( 3PP + 4PP \implies 5PP \)
2. if A and \( 3PP + 4PP \), then \( \delta(P, P_{n+1}) \leq \delta(P, P_n) \) for \( n = 0, 1, \ldots \)

where \( P \) is that \( P \) for which A holds.
Second Main Theorem

Proof.

- We saw that 5PP + alternating minimization implies Theorem 25.6.4.
- Combining 3PP and 4PP we have:

\[ Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \]  
\[ d(P, Q_0) - \delta(P, P_1) \geq d(P_1, Q_0) \]  (25.95)  
\[ \delta(P, P_1) + d(P, Q) \geq d(P, Q_1) \]  (25.96)  
\[ 3PP \]  
\[ \delta(P, P_1) + d(P, Q) \geq d(P, Q_1) \]  (25.97)  
\[ 4PP \]

- If we only consider \( Q_0 \) with \( d(P, Q_0) < \infty \) then \( \delta(P, P_1) < \infty \) since \( d(P_1, Q_0) \) is also finite (since \( d(P_1, Q_0) \leq d(P, Q_0) \) by \( Q_0 \xrightarrow{2} P_1 \)).

- So we can add the two above:

\[ d(P, Q_0) + d(P, Q) \geq d(P, Q_1) + d(P_1, Q_0) \]  (25.98)  
\[ \geq d(P, Q_1) + d(P_1, Q_1) \]  (25.99)

Further, if of both 3 and 4 points property hold, then if

\[ Q_n \xrightarrow{2} P_{n+1} \text{ in 3PP and } P_n \xrightarrow{1} Q_n \text{ in 4PP} \]

we get

\[ \delta(P, P_{n+1}) + d(P_{n+1}, Q_n) \leq d(P, Q_n) \leq \delta(P, P_n) + d(P, Q) \]  (25.100)

- This implies

\[ \delta(P, P_{n+1}) \leq \delta(P, P_n) + [d(P, Q) - d(P_{n+1}, Q_n)] \forall Q \]  (25.101)

so

\[ \delta(P, P_{n+1}) \leq \delta(P, P_n) + \left[ d(P, Q) - d(P_{n+1}, Q_n) \right] \leq 0 \]  (25.102)
Second Main Theorem

**Proof.**

- Implying that $\delta(P, P_{n+1}) \leq \delta(P, P_n)$

- Note, this shows that:
  \[
  \lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \tag{25.103}
  \]

- Ideally, we would like $d(P_0, Q) = d(P, Q)$
- True of course if $d() < \infty$ for all $P, Q$, but note that KL-divergence is not so.
- may depend on the starting value $P_0$, so in applications it is important to select a good starting value.

We will see later that if $\mathcal{P}$ and $\mathcal{Q}$ are convex and if $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ are measures on $(X, \mathcal{X})$ where $X$ is finite (e.g., discrete probability measures), and if we take $P_0$ to be such that $P_0(x) > 0$ if $\exists P \in \mathcal{P}, Q \in \mathcal{Q}$ s.t. $P(x)Q(x) > 0$, then

\[
\mathcal{P}_0 = \{P : D(P||Q) < +\infty\} \tag{25.104}
\]

is such that $d(\mathcal{P}_0, \mathcal{Q}) = d(\mathcal{P}, \mathcal{Q})$
Example

- Let $\mathcal{P}, \mathcal{Q}$ be closed convex subsets of a Hilbert space (normed space with a dot product s.t., every Cauchy sequence converges). Assume, e.g., $\mathbb{R}^n$
- Define $d(P, Q) = \|P - Q\|^2$ and $\delta(P, P') = \|P - P'\|^2$.
- This satisfies 3PP since Pythagorean theorem for right triangles, and that main angle will always be $\geq \pi/2$.

Thus, it satisfies 5PP.

This also satisfies 5PP sine angles at $Q_1$ are $\geq \pi/2$ (exercise: prove this).