## 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):** computing $R(D)$, alternating minimization
- **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 (but not this afternoon).
- 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.
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

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

The update is:

$$q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x, \hat{y})}} \quad (25.16)$$

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.
Computing $R(D)$

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

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

(25.16)

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

(25.17)

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

$$C(\hat{x}) = \sum_x p(x) e^{-\lambda d(x,\hat{x})}$$

$$\sum_{\hat{y}} q(\hat{y}) e^{-\lambda d(x,\hat{y})} = 1$$

(25.18)

- 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 $\lambda$.

Theorem 25.2.1

$\forall s > -\infty$, 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:

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

(25.16)

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

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

(25.17)
Computing $R(D)$

Theorem 25.2.1

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.

$$
R' = \frac{dR}{dD} \bigg|_{D_s} = s
$$

(25.18)

Proof.

Take derivatives and use the chain rule . . .

Pictorially,

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

$$
D = \sum_x p(x, \hat{x}) q(\hat{x}) e^{-\lambda d(x, \hat{x})} = D(\lambda, \{q(\hat{x})\})
$$

(25.16)

It can also be shown, moreover, that

$$
R = -(\lambda D + \sum_x p(x) \log \mu(x)) = R(\lambda, \{q(\hat{x})\})
$$

(25.17)

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

(25.18)

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 $R(D)$

- 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 (KKT conditions allow us to consider the case when $q(\hat{x}) > 0$ vs. $q(\hat{x}) \leq 0$).
- 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.

2 convex sets

- Consider the problem: we have two convex sets $A, B \subseteq \mathcal{R}^n$.
- We have a distance (e.g., Euclidean, or 2-norm) $d(a, b)$.
- Goal is to form:

$$d_{\text{min}} = \min_{a \in A, b \in B} d(a, b)$$  \hspace{1cm} (25.18)

- Consider the following algorithm:

1. Chose $a_0 \in A$ arbitrarily ;
2. for $n = 1 \ldots$ do
3. \hspace{1cm} Choose $b_n \in \text{argmin}_{b \in B} d(a_{n-1}, b)$ ;
4. \hspace{1cm} Choose $a_n \in \text{argmin}_{a \in A} d(a, b_n)$ ;
Alternating Minimization Between 2 convex sets

- In the slides that will soon follow, we will prove the conditions under which it will be the case that

\[
\lim_{n \to \infty} d(a_n, b_n) = d(a^*, b^*)
\] (25.18)

where

\[
(a^*, b^*) = \arg\min_{a,b} d(a, b)
\] (25.19)

and

\[
\lim_{n \to \infty} a_n = a^*, \quad \lim_{n \to \infty} b_n = b^*
\] (25.20)

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

---

Projections and I-maps

**Theorem 25.2.1**

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

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

\[
D(p(x)p(y|x)||p(x)r^*(y)) = \min_{r(y) \in \Delta} D(p(x)p(y|x)||p(x)r(y))
\] (25.18)

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

\[
\max_{r(x|y) \in \Delta^2} \sum_{x,y} p(x)p(y|x) \log \frac{r(x|y)}{p(x)} = \sum_{x,y} p(x)p(y|x) \log \frac{r^*(x|y)}{p(x)}
\] (25.19)
Another way of seeing these theorems is that we have:

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

(25.1)

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

(25.2)

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

(25.3)

Therefore, we have that for any \( r(y) \) and \( r(x|y) \), we have that:

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

(25.4)

\[
\sum_{x,y} p(x, y) \log \frac{r^*(x|y)}{r(x|y)} - \sum_{x,y} p(x, y) \log \frac{r(x|y)}{p(x)}
\]

(25.5)

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

(25.6)

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

(25.7)

\[
= D(r^*(x|y)||r(x|y))
\]

(25.8)

\[
\geq 0
\]

(25.9)

with equality when \( r^*(x|y) = r(x|y) \).
Therefore, we can write the rate-distortion function as

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

(25.10)

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

(25.11)

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

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

(25.12)

where

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

(25.13)

\[
B = \{ q(x, \hat{x}) : q(x, \hat{x}) = p(x)r(\hat{x}) \text{ for arbitrary } r(\hat{x}) \}
\]

(25.14)
Computing $R(D)$

- 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
  \[
  q(\hat{x}|x) = \frac{r(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_\hat{y} r(\hat{y})e^{-\lambda d(x,\hat{y})}}
  \]  
  (25.15)
- Once we have $q(\hat{x}) = q(\hat{x}|x)p(x)$, we find corresponding next $r(\hat{x})$ from the projection
  \[
  r(\hat{x}) = \sum_x p(x)q(\hat{x}|x)
  \]  
  (25.16)
- We repeat this alternating projection/minimization procedure until convergence.
- This will converge to $R(D)$ at $s$.

Computing Channel Capacity

- 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:
  \[
  C = \max_{q(x|y)} \max_{r(x)} \sum_{x,y} r(x)p(y|x) \log \frac{q(x|y)}{r(y)}
  \]  
  (25.17)
- We guess, starting $r(x)$ and then iterate the following two equations:
  \[
  q(x|y) = \frac{r(x)p(y|x)}{\sum_x r(x)p(y|x)}, \quad r(x) = \frac{\prod_y [q(x|y)p(y|x)]}{\sum_x \prod_y [q(x|y)p(y|x)]}
  \]  
  (25.18)
Alternating Minimization: Overall Idea

- Let \( P, Q \) 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 \( Q_n \in \text{argmin}_{Q \in \mathcal{Q}} D(P_n || Q) \).
- Then we have the following procedure:
  \[
  Q_n \in \text{argmin}_{Q \in \mathcal{Q}} D(P_n || Q) \tag{25.19}
  \]
  \[
  P_{n+1} \in \text{argmin}_{P \in \mathcal{P}} D(P || Q_n) \tag{25.20}
  \]
- Then the result we will get is that:
  \[
  D(P_n || Q_n) \rightarrow \inf_{(P,Q) \in (\mathcal{P}_0, \mathcal{Q})} D(P || Q) = D_{\text{min}} \tag{25.21}
  \]
  where \( \mathcal{P}_0 = \{P \in \mathcal{P} : D(P || Q_n) < \infty \text{ for some } n\} \) and \( P_n \rightarrow P^*, \quad Q_n \rightarrow Q^* \) sometimes as well, where \( D(P^*, Q^*) = D_{\text{min}} \).
- \( \mathcal{P}_0 \) are the entries of \( \mathcal{P} \) that we care about.

Alternating Minimization: Overall Idea

- This process has a geometric flavor, since it corresponds to alternating “projections” based on treating KL as a generalized “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 \( d : \mathcal{P} \times \mathcal{Q} \rightarrow \mathbb{R} \cup \{+\infty\} \) which need not be KL-divergence.
Distance/Metric/Etc.

It is typical to be formal about such terms (recall from Lecture 3).

- Let $X$ be a set. A function $d : X \times X \to \mathbb{R}$ is called a **distance** on $X$ if, $\forall x, y \in X$, we have $d(x, y) \geq 0$ (non-negativity), $d(x, y) = d(y, x)$ (symmetry), and $d(x, x) = 0$ (reflexivity).
- Let $X$ be a set. A function $d : X \times X \to \mathbb{R}$ is called a **quasi-distance** on $X$ if it is non-negative and reflexive.
- Let $X$ be a set. A function $d : X \times X \to \mathbb{R}$ is called a **semi-metric** on $X$ if it is non-negative, symmetric, reflexive, and if $d(x, y) \leq d(x, z) + d(z, y)$ for all $x, y, z \in X$ (triangular inequality).
- Let $X$ be a set. A function $d : X \times X \to \mathbb{R}$ is called a **metric** on $X$ if it is a semi-metric and if $d(x, y) = 0$ iff $x = y$ (identity of indiscernibles).
- Hence, the KL-divergence is like a quasi-distance that also satisfies identity of indiscernibles. A reasonable name for this is a “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 $P \xrightarrow{1} Q'$ 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 $Q \xrightarrow{2} P'$, indicating we minimize over $P$, holding the 2nd argument $Q$ fixed.
- Sequences obtained by alternating minimization $\{(P_n, Q_n)\}_{n=0}^{\infty}$ as:
  $$P_0 \xrightarrow{1} Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \xrightarrow{2} P_2 \xrightarrow{1} Q_2 \xrightarrow{2} P_3 \xrightarrow{1} Q_3 \xrightarrow{2} \cdots$$  
  (25.22)
  where we start arbitrarily with $P_0$.
- Goal: sufficient conditions for the convergence of the alternating minimization procedure.
**Five Points Property**

**Definition 25.4.1 (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 at $P \in \mathcal{P}$ if: $\forall Q \in \mathcal{Q}, \forall Q_0 \in \mathcal{Q}$, we have:

$$d(P, Q) + d(P, Q_0) \geq d(P, Q_1) + d(P_1, Q_1)$$  \hspace{1cm} (25.23)

whenever $Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1$. We say $d(\cdot, \cdot)$ 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 ☺).
- **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

\[
\lim_{n \to \infty} d(P_n, Q_n) = \inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} d(P, Q) = d_{\text{min}} \tag{25.24}
\]

as long as

\[
d_{\text{min}} = \inf_{P \in \mathcal{P}_0, Q \in \mathcal{Q}} d(P, Q) \tag{25.25}
\]

where

\[
\mathcal{P}_0 = \{ P : P \in \mathcal{P}, d(P, Q_n) < \infty \text{ for some } n \} \tag{25.26}
\]

- Note, \( \mathcal{P}_0 \) does depend on the sequence, and \( \mathcal{P}_0 = \mathcal{P} \) if \( d \) is finite valued.

Definitions

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

\[
d(A, B) \triangleq \inf_{P \in A, Q \in B} d(P, Q) \tag{25.27}
\]

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

**Lemma 25.4.2**

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

\[
d(P_n, Q_n) \geq d(\mathcal{P}_0, \mathcal{Q}) \quad \forall n \tag{25.28}
\]

**Proof.**

Obvious via definitions.

Our goal is to first find when \( \lim_{n \to \infty} d(P_n, Q_n) = d(\mathcal{P}_0, \mathcal{Q}) \).
Recall that \( \limsup \) is different than \( \lim \),

\[
\limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S \tag{25.29}
\]

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.

Also,

\[
\liminf_{n \to \infty} a_n \triangleq \sup_{n > 0} \left( \inf_{k > n} a_k \right) \tag{25.30}
\]

so \( \liminf \) asks for supremum convergence in the local minima.

---

**Key Lemma**

**Lemma 25.4.3**

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:

\[
c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots. \tag{25.31}
\]

And also assume that

\[
\limsup_{n \to \infty} b_n > -\infty, \quad \text{and} \quad \exists n_0 \text{ s.t. } b_{n_0} < \infty. \tag{25.32}
\]

Then

\[
\liminf_{n \to \infty} a_n \leq c \tag{25.33}
\]

Also, if in addition, we assume that

\[
\sum_{n=0}^\infty (c - a_n)^+ < \infty \quad \text{then} \quad \sum_{n=n_0+1}^\infty |a_n - c| < \infty \tag{25.34}
\]

and as a result

\[
\lim_{n \to \infty} a_n = c \tag{25.35}
\]
**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’ have that \( \lim \inf_{n \to \infty} a_n > c \) 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.**

- Next if \( b_{n_0} < \infty \) for some \( n_0 \), then since \( c \) is finite, and since
  \[
  c + b_{n-1} \geq b_n + a_n, \tag{25.36}
  \]
  then we have \( a_n < \infty, b_n < \infty, \forall n > n_0 \).
- Thus, \( a_n - c \leq b_{n-1} - b_n \) for \( n > n_0 \), giving:
  \[
  \sum_{n=n_0+1}^{n} (a_n - c) \leq \sum_{n=n_0+1}^{n} (b_{n-1} - b_n) = b_{n_0} - b_n \quad \forall n > n_0 \tag{25.37}
  \]
- Since \( \limsup_{n \to \infty} b_n > -\infty \) (by assumption), and \( b_n < \infty \) for \( n > n_0 \), and if \( \sum_{n=0}^{\infty} (c - a_n)^+ < \infty \), we have that (exercise)
  \( \lim_{n \to \infty} b_n - b_{n_0} > -\infty \), or \( \lim_{n \to \infty} b_{n_0} - b_n < \infty \), meaning that it has a limit and \( \sum_{n=0}^{\infty} (c - a_n)^+ < \infty \).
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 \( a = a^+ - a^- = c_+ < \infty \) and if \( a^+ = c_+ < \infty \), then \( |a| = a^+ + a^- = -c_+ < \infty \).
- Then when \( \sum_{n=n_0+1}^{\infty} |a_n-c| < \infty \), this means that \( \lim_{n \to \infty} a_n = c \).

Restated, since \( \sum_{n=n_0+1}^{N} (c-a_n)^+ < \infty \), this means that series \( S_N = \sum_{n=n_0+1}^{N} (c-a_n)^+ \) has a limit, \( N \geq n_0 + 1 \), and that also \( R_N = \sum_{n=n_0+1}^{N} (a_n-c) \) also has a limit (\( \lim_{N \to \infty} R_N \) exists in the extended reals).

Also, if the limit is finite, then we have

\[
\sum_{n=n_0+1}^{\infty} (a_n-c) < \infty \Rightarrow \sum_{n=n_0+1}^{\infty} (a_n-c)^+ < \infty \Rightarrow \sum_{n=n_0+1}^{\infty} (c-a_n)^- < \infty
\]

This and \( \sum_{n=n_0+1}^{\infty} (a_n-c)^+ < \infty \) means \( \sum_{n=n_0+1}^{\infty} (c-a_n)^- + (c-a_n)^+ < \infty \)

Implying that \( \sum_{n=n_0+1}^{\infty} |c-a_n| < \infty \) or \( \lim_{n \to \infty} a_n = c \).
1st Main theorem

**Theorem 25.4.4**

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

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

(25.38)

**Note:** no minimization done here, only 5PP condition on the sequences. Then if either: A) \( \forall P \in \mathcal{P}_0 \); or B) for some \( P \in \mathcal{P}_0 \) s.t. \( d(P, Q) = d(\mathcal{P}_0, Q) \), we have:

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

(25.39)

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

\[
\sum_{n=0}^{\infty} (d(P_n, Q_n) - d(\mathcal{P}_0, Q)) < \infty
\]

(25.40)

**Proof of 1st main theorem**

**Proof.**

- If \( \mathcal{P}_0 = \emptyset \) then, for all \( n \geq 1 \), we have
  
  \[
d(P_n, Q_n) = d(\mathcal{P}_0, Q) = \inf_{P \in \mathcal{P}_0, Q \in \mathcal{Q}} d(P, Q) = \inf \emptyset = \infty
\]
  
  (25.41)

  so theorem is true in this case (l.h.s. holds by definition of \( \mathcal{P}_0 \)).

- Suppose \( \mathcal{P}_0 \neq \emptyset \) and that Eq (25.38) holds for some \( P \in \mathcal{P}_0 \).

- Then, lemma 25.4.3 \( (c + b_{n-1} \geq b_n + a_n, \ n = 1, 2, \ldots) \) with

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

  (25.42)

- Why? Since \( P \in \mathcal{P}_0 \), we have both \( \exists n_0 \) s.t., \( b_{n_0} < \infty \) and also \( c < \infty \) (all by the def of \( \mathcal{P}_0 \)) implying \( a_n < \infty \) for \( n \geq n_0 \).

- Also, \( \lim \sup_{n \to \infty} b_n > -\infty \) since Eq (25.38) 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 25.4.3 holds here and from it we get:
  - Under A: \( \forall P \in \mathcal{P}_0 \), we have
    \[
    \liminf_{n \to \infty} a_n = \liminf_{n \to \infty} d(P_n, Q_n) \leq c = d(P, Q) < \infty \quad (25.43)
    \]
    - Thus, \( \forall P \in \mathcal{P}_0 \), we have that \( d(P_n, Q_n) \) “converges” to a finite value (not \( \infty \)) since \( d > -\infty \) (since it holds for all \( P \in \mathcal{P}_0 \), we have \( \liminf_{n \to \infty} d(P_n, Q_n) \leq d(\mathcal{P}_0, Q) < \infty \)).
  - Recall, \( d(P_n, Q_n) \geq d(\mathcal{P}_0, Q) \) for \( n = 0, 1, \ldots \) for any sequence \( \{P_n, Q_n\}\) for \( n = 0, 1, \ldots \)
  - Also, let \( P = P_{n-1} \) in Eq (25.38), so we get:
    \[
    d(P_{n-1}, Q) + d(P_{n-1}, Q_{n-1}) \geq d(P_{n-1}, Q_n) + d(P_n, Q_n) \quad (25.44)
    \]

This implies that

\[
\begin{align*}
    d(P_{n-1}, Q_{n-1}) & \geq \underbrace{d(P_{n-1}, Q_n) - d(P_{n-1}, Q) + d(P_n, Q_n)}_{\geq 0} \\
& \geq d(P_n, Q_n)
\end{align*}
\quad (25.45)
\]

- 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 \( d(\mathcal{P}_0, Q) \) means that
  \[
  \lim_{n \to \infty} d(P_n, Q_n) = d(\mathcal{P}_0, Q) \quad (25.46)
  \]
Proof of 1st main theorem

Proof.

Next, under B (for some \( P \in \mathcal{P}_0 \) s.t. \( d(P, Q) = d(P_0, Q) \)), we have that

\[
d(P_n, Q_n) \geq d(P_0, Q) = d(P, Q) \tag{25.47}
\]

which follows since, as mentioned earlier, \( d(P_n, Q_n) \geq d(P_0, Q) \) for \( n = 0, 1, \ldots \) for any sequence \((\{P_n, Q_n\})_{n=0}^{\infty}\).

This means,

\[
a_n = d(P_n, Q_n) \geq c = d(P, Q) \tag{25.48}
\]

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

From lemma 25.4.3, this gives \( \lim_{n \to \infty} a_n = c \) or

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P, Q) = d(P_0, Q) \tag{25.49}
\]

And since (as shown earlier)

\[
\sum_{n=n_1}^{\infty} (a_n - c) < \infty \tag{25.50}
\]

we have

\[
\sum_{n=n_1}^{\infty} (d(P_n, Q_n) - d(P_0, Q)) < \infty \tag{25.51}
\]

\[
(25.52) \quad (25.53)
\]
Consider next sequences \( \{(P_n, Q_n)\}_{n=0}^{\infty} \) constructed by alternating minimization with arbitrary starting point \( P_0 \in \mathcal{P} \)

\[
P_0 \xrightarrow{1} Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \xrightarrow{2} P_2 \xrightarrow{1} Q_2 \xrightarrow{2} P_3 \xrightarrow{1} Q_3 \xrightarrow{2} \ldots \quad (25.54)
\]

Then we have that:

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

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 25.4.4 met at \( P \in \mathcal{P}_0 \)

That is, for \( n = 1 \) we have \( Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \) so

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

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

Thus, this also certainly holds for \( Q_0 \) such that \( P_0 \xrightarrow{1} Q_0 \).

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

For \( n = 2 \) we have \( Q_1 \xrightarrow{2} P_2 \xrightarrow{1} Q_2 \) so

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

so also true for \( Q_1 \) such that \( P_1 \xrightarrow{1} Q_1 \).

Same for \( n > 2 \), etc.
Sequences

- So Theorem 25.4.4 holds in this case (i.e., \( \lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \)).
- 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 25.4.5 (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 \( \forall Q_0 \)

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

(25.58)

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

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

\( P_1 \in \text{argmin}_{P \in P} d(P, Q_0) \)

Four Points Property (4PP)

**Definition 25.4.6 (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

\[ d(P, Q_1) \leq \delta(P, P_1) + d(P, Q) \] whenever \( P_1 \xrightarrow{1} Q_1 \)  \hspace{1cm} (25.59)
**Four Points Property (4PP)**

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

\[ Q_1 \in \arg\min_{Q \in Q} d(P_1, Q) \]

**Second Main Theorem**

**Theorem 25.4.7**

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

\[ \lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \quad (25.60) \]

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. \( 3PP + 4PP \Rightarrow 5PP \)
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 25.4.4.
- Combining 3PP and 4PP we have:

\[
Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \tag{25.61}
\]
\[
d(P, Q_0) - \delta(P, P_1) \geq d(P_1, Q_0) \quad \text{3PP} \tag{25.62}
\]
\[
\delta(P, P_1) + d(P, Q) \geq d(P, Q_1) \quad \text{4PP} \tag{25.63}
\]

- 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:

\[
d(P, Q_0) + d(P, Q) \geq d(P, Q_1) + d(P_1, Q_0) \tag{25.64}
\]
\[
\geq d(P, Q_1) + d(P_1, Q_1) \tag{25.65}
\]

Further, if of both 3 and 4 points property hold, then if \( Q_n \xrightarrow{2} P_{n+1} \) in 3PP and \( P_n \xrightarrow{1} Q_n \) in 4PP

we get

\[
\delta(P, P_{n+1}) + d(P_{n+1}, Q_n) \leq d(P, Q_n) \leq \delta(P, P_n) + d(P, Q) \tag{25.66}
\]

This implies

\[
\delta(P, P_{n+1}) \leq \delta(P, P_n) + [d(P, Q) - d(P_{n+1}, Q_n)] \forall Q \tag{25.67}
\]

so

\[
\delta(P, P_{n+1}) \leq \delta(P, P_n) + \sum_0 \left[ d(P, Q) - d(P_{n+1}, Q_n) \right] \tag{25.68}
\]
Second Main Theorem

Proof.

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

- Note, this shows that:
  \[
  \lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \tag{25.69}
  \]

- 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.

It turns out 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 $P_0(x) > 0$ and if $\exists P \in \mathcal{P}, Q \in \mathcal{Q}$ s.t. $P(x)Q(x) > 0$, then it is the case that

\[
\mathcal{P}_0 = \{ P : D(P||Q) < +\infty \} \tag{25.70}
\]

has the property that $d(\mathcal{P}_0, Q) = d(\mathcal{P}, Q)$
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 $\geq \pi/2$.

Example: Squared Euclidean-Distance

- Let $\mathcal{P}$ and $\mathcal{Q}$ be two non-empty closed convex subsets of $\mathbb{R}^N$.
- Let $d(P, Q) = \|P - Q\|^2$ and $\delta(P, P') = \|P - P'\|^2$
- Goal is to find $\min_{P \in \mathcal{P}, P \in \mathcal{Q}} d(P, Q)$
- Let $P, Q_0$ be given, and $P_1 = \arg\min_{P' \in \mathcal{P}} d(P', Q_0)$. Three points property states that, for all $P$, $d(P, Q_0) \geq \delta(P, P_1) + d(P_1, Q_0)$.
- This follows since:

\[
\|P - Q_0\|^2 = \|P - P_1 + P_1 - Q_0\|^2 \quad \quad \quad (25.71)
= \|P - P_1\|^2 + \|P_1 - Q_0\|^2 + 2 \langle P - P_1, P_1 - Q_0 \rangle \quad \quad \quad (25.72)
\]

- This gives the result since $\langle P - P_1, P_1 - Q_0 \rangle \geq 0$
Example: Squared Euclidean-Distance

- Geometry of why $\langle P - P_1, P_1 - Q_0 \rangle \geq 0$
- Note that $P_1$ is the orthogonal projection of $Q_0$ onto (convex set) $P$

Thus, since squared Euclidean-Distance satisfies both 3PP and 4PP, it satisfies 5PP.
Convex Sets

- So far, we have not required the sets to be convex (although of course the example in the squared-euclidean case, was).
- Let \((X, \mathcal{X})\) be a measurable space with sets of finite measures \(P, Q\) on \((X, \mathcal{X})\). That is, for all \(E \in \mathcal{X}\), \(P(E) < \infty\) for all \(P \in \mathcal{P}\).
- Let \(\mathcal{P}, \mathcal{Q}\) be convex (e.g., simplex or convex subsets thereof).
- Define

\[
d(P, Q) = D(P||Q) = \begin{cases} 
\int \log p dP & \text{if } P \ll Q \\
\infty & \text{if } P \not\ll Q
\end{cases}
\]

(25.73)

where \(p = \frac{dP}{dQ}\) is the Radon-Nikodym derivative (more later).
- Note, \(P \ll Q\) is spoken as "\(Q\) dominates \(P\)" meaning for all \(E\) such that \(Q(E) = 0\), \(P(E) = 0\) (when \(Q\) becomes zero, it forces \(P\) to also be zero).

\[
p = \frac{dP}{dQ}\text{ is the Radon-Nikodym derivative, meaning that } \forall E \in \mathcal{X}, \text{ we have that } p \text{ is defined so that}
\]

\[
p(E) = \int_E \left[\frac{dP}{dQ}\right] dQ
\]

(25.74)

i.e., we see \(\frac{dP}{dQ}\) as a function that maps one measure to the other.
- Thus, if \(P \ll Q\), then

\[
D(P||Q) = \int \log \left[\frac{dP}{dQ}\right] dP = \int p(x) \log \frac{p(x)}{q(x)} dx
\]

(25.75)

This basically means that \(\{q(x) = 0\} \Rightarrow \{p(x) = 0\}\).
- And if \(P \not\ll Q\), then, what happens? Could have \(q(x) = 0\) without \(p(x) = 0\).
For $P, P' \in \mathcal{P}$, we have the generalized KL as follows:

$$\delta(P, P') \triangleq D(P||P') + P'(X) - P(X) \geq 0 \quad (25.76)$$

- So if the measures are probability measures then $P'(X) = P(X) = 1$, and we get back standard KL-divergence.
- If not probability measures, $D(P||P')$ could go negative, but not $\delta(\cdot, \cdot)$.

---

### Theorem 25.5.1

The 3PP holds in this setting. I.e., let $\mathcal{P}$ be convex sets of measures, $Q_0$ be another measure on $(X, \mathcal{X})$. Then if $Q_0 \overset{1}{\to} P_1$, then we have

$$D(P||P_1) + P_1(X) - P(X) + D(P_1||Q_0) \leq D(P||Q_0) \quad (25.77)$$

#### Proof.

- By definition, $D(P_1||Q_0) = D(P||Q_0) < \infty$.
- Assume $D(P||Q_0) < \infty$ since otherwise we get the inequality immediately. Next, define

$$p_1 = \frac{dP_1}{dQ_0}, \text{ and } p = \frac{dP}{dQ_0} \quad (25.78)$$

so that $P_1 = \int p_1 dQ_0$ and $P = \int p dQ_0$
Proof.

Form \( P_\alpha = (1 - \alpha)P + \alpha P_1 \) as a convex combination of \( P \) and \( P_1 \)

and then \( f(\alpha) \triangleq D(P_\alpha||Q_0) \) so

\[
f(1) = D(P_1||Q_0) < D(P_\alpha||Q_0) = f(\alpha)
\] (25.79)

since \( Q_0 \stackrel{2}{\rightarrow} P_1 \).

Thus,

\[
0 \geq \frac{f(1) - f(\alpha)}{1 - \alpha} = \frac{1}{1 - \alpha} \left[ \int dP_1 \log \frac{dP_1}{dQ_0} - \int dP_\alpha \log \frac{dP_\alpha}{dQ_0} \right]
\] (25.80)

\[
= \frac{1}{1 - \alpha} \left[ \int p_1 dQ_0 \log \frac{dP_1}{dQ_0} - \int p_\alpha dQ_0 \log \frac{dP_\alpha}{dQ_0} \right]
\] (25.81)

\[
= \frac{1}{1 - \alpha} \left[ \int (p_1 \log p_1 - p_\alpha \log p_\alpha) dQ_0 \right]
\] (25.82)

From this, the results follow using definitions of \( p_1 \) and \( p_\alpha \)...

By certain technical reasons, we can exchange the limit (or really derivative) and the integral to get:

\[
0 \geq \int \frac{d}{d\alpha} (p_\alpha \log p_\alpha)|_{\alpha=1}dQ_0 = \int (1 + \log p_1)(p_1 - p)dQ_0
\] (25.84)

\[
= \int p_1 dQ_0 - pdQ_0 + p_1 \log p_1 dQ_0 - p \log p_1 dQ_0
\] (25.85)

It can be shown that this is non-increasing as \( \alpha \uparrow 1 \).

Moreover, since \( p_\alpha \log p_\alpha \) is convex, we can find the max of the above by taking derivatives, but the derivative is what happens when \( \alpha \uparrow 1 \). That is,

\[
\lim_{\alpha \uparrow 1} (\text{quantity}) = \frac{d}{d\alpha} (\text{quantity}) \bigg|_{\alpha=1}
\] (25.83)
4PP KL

Theorem 25.5.2

Let \( Q \) be a convex set of measures and let \( P_1 \) be a measure on \((X, \mathcal{X})\), then \( P_1 \rightarrow Q_1 \) yields

\[
D(P || Q_1) \leq D(P || P_1) + P_1(X) - P(X) + D(P || Q) \tag{25.86}
\]
\[
d(P, Q_1) \leq \delta(P, P_1) + d(P, Q) \tag{25.87}
\]

for all \( P \) on \((X, \mathcal{X})\) and for all \( Q \in Q \).

- So 4PP holds as well (proof skipped) so 5PP holds in this case as we..

A main theorem

Theorem 25.5.3

Let \( \mathcal{P}, \mathcal{Q} \) be convex sets of measures on \((X, \mathcal{X})\) with \( \{(P_n, Q_n)\}_{n=0}^{\infty} \) be sequences from \( \mathcal{P}, \mathcal{Q} \) obtained from alternating minimization of \( d(P, Q) = D(P || Q) \), starting from \( P_0 \in \mathcal{P} \), then we have that

\[
\lim_{n \to \infty} D(P_n || Q_n) = D(\mathcal{P}_0 || \mathcal{Q}) \tag{25.88}
\]

where

\[
\mathcal{P}_0 = \{ P : D(P || Q_n) < \infty \text{ for some } n \} \tag{25.89}
\]

Also, if \( X \) is a finite set (+ a few other minor technical conditions) then
\( P_n \rightarrow P^* \) where \( D(P^* || Q) = D(\mathcal{P}_0 || \mathcal{Q}) \).
A main theorem (continued)

- Moreover, if $X$ is finite, and $P_0$ is positive for $x \in X$ such that $\exists P \in \mathcal{P}, Q \in \mathcal{Q}$ with $P(x)Q(x) > 0$ (simultaneously positive on $X$), then $\mathcal{P}_0 = \{P : D(P || Q) < \infty\}$ so that $D(\mathcal{P}_0 || Q) = D(\mathcal{P} || Q)$ and we get the sequence independent minimum (i.e., $\mathcal{P}_0$ no longer depends on the sequence).
- Bottom line: when you implement this, make sure to initialize each distribution to strictly positive values.
- How to know when to stop in practice?
- Can we bound $D(P_n || Q_n) - D(P_0 || Q)$?
- If $D(P || Q) = D(P_0 || Q)$ then from 5PP we get
  \[
  D(\mathcal{P}_0 || Q) + D(P || Q_{n-1}) \geq D(P || Q_n) + D(P_n || Q_n) \tag{25.90}
  \]

This implies that

\[
D(P_n || Q_n) - D(P_0 || Q) \leq D(P || Q_{n-1}) - D(P || Q_n) \tag{25.91}
\]

\[
= \int \log \frac{dQ_n}{dQ_{n-1}} dP \tag{25.92}
\]

\[
\leq \log \sup_x \frac{dQ_n(x)}{dQ_{n-1}(x)} \to 0 \tag{25.93}
\]

if the sequence has a limit (i.e., is convergent) so that it does not change from iteration to iteration.
- Thus we can upper bound how close we are to convergence by looking a ratios of successive measures.
Practical Considerations

- In general, one is not guaranteed that each minimization will be easy but for alternating minimization on KL-divergence, it is (as we have already seen for both rate-distortion theory and channel capacity).
- Also, in machine learning, the “measure propagation” algorithm, and the “label propagation” algorithm for semi-supervised learning can be seen to satisfy all of the above properties, and so the convergence holds there as well.

New topic

- Next time, we’ll start looking at complexity measures in general, and begin our discussion of alternative complexity measures (including Kolmogorov or algorithmic complexity).