Outstanding Reading

- Read all chapters assigned from IT-I (EE514, Winter 2012).
- Read chapter 8 in the book.
- Read chapter 9 in the book.
- Read chapter 10 in the book (chapter on rate distortion theory).
Additional Reading on Rate-Distortion Theory

- “Information Geometry and Alternating Minimization Procedures”, Csiszár & Tusnády, 1983
Please do use our discussion board (https://catalyst.uw.edu/gopost/board/bilmes/27386/) for all questions, comments, so that all will benefit from them being answered.
On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
On Final Presentations

- Your task is to give a 15-20 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.
On Final Presentations

- Your task is to give a 15-20 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).
On Final Presentations

- Your task is to give a 15-20 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 papers must have been published in the last 10 years (so no old or classic papers).
On Final Presentations

- Your task is to give a 15-20 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 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.
On Final Presentations

- Your task is to give a 15-20 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 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.
- 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.
On Final Presentations

- Your task is to give a 15-20 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 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.
- 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 dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- Wed, May 2nd: Candidate proposed papers submitted. Include short at most 1-page writeup: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.
Final Presentation Milestones

All submissions done in PDF file format via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- **Wed, May 2nd:** Candidate proposed papers submitted. Include short at most 1-page writeup: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.

- **Friday, May 11th:** Updated list of proposed papers decided, based on feedback. Updated writeup.

- **Friday, May 18th:** short writeup on more details of how you will present the ideas in a simple fashion.

- **Friday, May 25th:** updated short writeup on more details of how you will present the ideas in a simple fashion.

- **Final presentations:** Monday, June 4th in the afternoon late/evening (currently scheduled for 8:30am but that is too early). What to turn in: your slides and a short at most 4 page summary of the papers.
Final Presentation Milestones

All submissions done in PDF file format via our dropbox
(https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- Wed, May 2nd: Candidate proposed papers submitted. Include short at most 1-page writeup: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.
- Friday, May 11th: Updated list of proposed papers decided, based on feedback. Updated writeup.
- Friday, May 18th: short writeup on more details of how you will present the ideas in a simple fashion.
Final Presentation Milestones

All submissions done in PDF file format via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- Wed, May 2nd: Candidate proposed papers submitted. Include short at most 1-page writeup: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.

- Friday, May 11th: Updated list of proposed papers decided, based on feedback. Updated writeup.

- Friday, May 18th: short writeup on more details of how you will present the ideas in a simple fashion.

- Friday, May 25th: updated short writeup on more details of how you will present the ideas in a simple fashion.
Final Presentation Milestones

All submissions done in PDF file format via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- Wed, May 2nd: Candidate proposed papers submitted. Include short at most 1-page writeup: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.

- Friday, May 11th: Updated list of proposed papers decided, based on feedback. Updated writeup.

- Friday, May 18th: short writeup on more details of how you will present the ideas in a simple fashion.

- Friday, May 25th: updated short writeup on more details of how you will present the ideas in a simple fashion.

- Final presentations: Monday, June 4th in the afternoon late/evening (currently scheduled for 8:30am but that is too early). What to turn in: your slides and a short at most 4 page summary of the papers.
Computing $R(D)$

Can restate this problem as:

$$R(D) = \min_{q(\hat{x}|x)} I(X; \hat{X}) \quad (1)$$

s.t. $q(\hat{x}|x) \geq 0 \ \forall \hat{x}, x \quad (2)$

$$\sum_{\hat{x}} q(\hat{x}|x) = 1 \ \forall x \quad (3)$$

$$\sum_{\hat{x},x} q(\hat{x}|x)p(x)d(x, \hat{x}) = D \quad (4)$$

where

$$I(X; \hat{X}) = \sum_{x,\hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{q(\hat{x})} \quad (5)$$

and $q(\hat{x}) = \sum_{x} p(x)q(\hat{x}|x) \quad (6)$
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.
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 quasi-distance \( d(\cdot, \cdot) \) and a constraint set \( \mathcal{P} \), and a vector \( \hat{x} \notin \mathcal{P} \).
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 quasi-distance \( d(\cdot, \cdot) \) and a constraint set \( P \), and a vector \( \hat{x} \notin P \).

- We want to find the member of \( P \) that is closest to \( \hat{x} \) where closeness is measured via \( d(\cdot, \cdot) \), i.e.,

\[
x^* \in \arg\min_{x \in P} d(\hat{x}, x)
\]  

\( x^* \in \arg\min_{x \in P} d(\hat{x}, x) \quad (7) \)
Distance, Divergence, and Quasi-Distance

We are going to be defining distance-like functions of the form:
\[ d : P \times Q \rightarrow \mathbb{R} \cup \{+\infty\} \]
where \( P \) and \( Q \) are sets.

Note that the terms distance, metric, etc. have specific meanings.
\[ d : P \times P \rightarrow \mathbb{R} \] is a distance if for all \( x, y \in P \),
\[ d(x, y) \geq 0 \] (non-negative),
\[ d(x, y) = d(y, x) \] (symmetric), and
\[ d(x, x) = 0 \] (reflexivity).

A quasi-distance is \( d : P \times P \rightarrow \mathbb{R} \) that is both non-negative and
\[ d(x, x) = 0 \] for all \( x \in P \) (note that this is not iff, and this is Deza's
definition in "Encyclopedia of Distances," but other definitions of
quasi-distance may be found).

By divergence, we might also allow the sets to be different.
Some people will really care about this (and if you accidently call
something a distance, they will reject your paper).
Thus, a warning: make sure you are clear about what kind of object
it is, and for the purposes of any given publication, if you call it
something, define it at the same time.
Distance, Divergence, and Quasi-Distance

- We are going to be defining distance-like functions of the form: 
  \[ d : \mathcal{P} \times \mathcal{Q} \rightarrow \mathbb{R} \cup \{+\infty\} \] 
  where \( \mathcal{P} \) and \( \mathcal{Q} \) are sets.
- Note that the terms distance, metric, etc. have specific meanings.
Distance, Divergence, and Quasi-Distance

- We are going to be defining distance-like functions of the form: $d : \mathcal{P} \times \mathcal{Q} \rightarrow \mathbb{R} \cup \{+\infty\}$ where $\mathcal{P}$ and $\mathcal{Q}$ are sets.
- Note that the terms distance, metric, etc. have specific meanings.
- $d : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$ is a distance if for all $x, y \in \mathcal{P}$, $d(x, y) \geq 0$ (non-negative), $d(x, y) = d(y, x)$ (symmetric), and $d(x, x) = 0$ (reflexivity).

A quasi-distance is $d : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$ that is both non-negative and $d(x, x) = 0$ for all $x \in \mathcal{P}$ (note that this is not iff, and this is Deza’s definition in “Encyclopedia of Distances,” but other definitions of quasi-distance may be found).

By divergence, we might also allow the sets to be different. Some people will really care about this (and if you accidently call something a distance, they will reject your paper). Thus, a warning: make sure you are clear about what kind of object it is, and for the purposes of any given publication, if you call it something, define it at the same time.
Distance, Divergence, and Quasi-Distance

- We are going to be defining distance-like functions of the form: 
  
  \[ d : P \times Q \rightarrow \mathbb{R} \cup \{+\infty\} \] 
  
  where \( P \) and \( Q \) are sets.

- Note that the terms distance, metric, etc. have specific meanings.

- \( d : P \times P \rightarrow \mathbb{R} \) is a distance if for all \( x, y \in P \), 
  \[ d(x, y) \geq 0 \] 
  (non-negative), 
  \[ d(x, y) = d(y, x) \] 
  (symmetric), and 
  \[ d(x, x) = 0 \] 
  (reflexivity).

- A quasi-distance is \( d : P \times P \rightarrow \mathbb{R} \) that is both non-negative and 
  \[ d(x, x) = 0 \] 
  for all \( x \in P \) (note that this is not iff, and this is Deza’s 
  definition in “Encyclopedia of Distances,” but other definitions of 
  quasi-distance may be found).
Distance, Divergence, and Quasi-Distance

- We are going to be defining distance-like functions of the form: 
  \( d : P \times Q \rightarrow \mathbb{R} \cup \{+\infty\} \) where \( P \) and \( Q \) are sets.
- Note that the terms distance, metric, etc. have specific meanings.
- \( d : P \times P \rightarrow \mathbb{R} \) is a distance if for all \( x, y \in P \), \( d(x, y) \geq 0 \) (non-negative), \( d(x, y) = d(y, x) \) (symmetric), and \( d(x, x) = 0 \) (reflexivity).
- A quasi-distance is \( d : P \times P \rightarrow \mathbb{R} \) that is both non-negative and \( d(x, x) = 0 \) for all \( x \in P \) (note that this is not iff, and this is Deza’s definition in “Encyclopedia of Distances,” but other definitions of quasi-distance may be found).
- By divergence, we might also allow the sets to be different.
We are going to be defining distance-like functions of the form:
\[ d : \mathcal{P} \times \mathcal{Q} \rightarrow \mathbb{R} \cup \{+\infty\} \]
where \( \mathcal{P} \) and \( \mathcal{Q} \) are sets.

Note that the terms distance, metric, etc. have specific meanings.

\[ d : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R} \]
is a distance if for all \( x, y \in \mathcal{P} \),
\[ d(x, y) \geq 0 \] (non-negative),
\[ d(x, y) = d(y, x) \] (symmetric), and
\[ d(x, x) = 0 \] (reflexivity).

A quasi-distance is \( d : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R} \)
that is both non-negative and
\[ d(x, x) = 0 \] for all \( x \in \mathcal{P} \)
(note that this is not iff, and this is Deza’s
definition in “Encyclopedia of Distances,” but other definitions of
quasi-distance may be found).

By divergence, we might also allow the sets to be different.

Some people will really care about this (and if you accidently call
something a distance, they will reject your paper).
Distance, Divergence, and Quasi-Distance

- We are going to be defining distance-like functions of the form:
  \[ d : \mathcal{P} \times \mathcal{Q} \rightarrow \mathbb{R} \cup \{ +\infty \} \]
  where \( \mathcal{P} \) and \( \mathcal{Q} \) are sets.

- Note that the terms distance, metric, etc. have specific meanings.

- \( d : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R} \) is a distance if for all \( x, y \in \mathcal{P} \),
  \( d(x, y) \geq 0 \) (non-negative), \( d(x, y) = d(y, x) \) (symmetric), and \( d(x, x) = 0 \) (reflexivity).

- A quasi-distance is \( d : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R} \) that is both non-negative and
  \( d(x, x) = 0 \) for all \( x \in \mathcal{P} \) (note that this is not iff, and this is Deza’s
  definition in “Encyclopedia of Distances,” but other definitions of
  quasi-distance may be found).

- By divergence, we might also allow the sets to be different.

- Some people will really care about this (and if you accidently call
  something a distance, they will reject your paper).

- Thus, a warning: make sure you are clear about what kind of object
  it is, and for the purposes of any given publication, if you call it
  something, define it at the same time.
Computing $R(D)$

- Both inequality (Eq.(2)) and equality constraints (Eqs.(3) & (4)).
Computing $R(D)$

- Both inequality (Eq. (2)) and equality constraints (Eqs. (3) & (4)).
- 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: intuitively, we know can only make $R(D)$ smaller by making $D$ larger.
Computing $R(D)$

- Both inequality (Eq.(2)) and equality constraints (Eqs.(3)&(4)).
- 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?
Computing $R(D)$

- Both inequality (Eq.(2)) and equality constraints (Eqs.(3) & (4)).
- 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: intuitively, we know can only make $R(D)$ smaller by making $D$ larger.
Computing $R(D)$

- Both inequality (Eq.(2)) and equality constraints (Eqs.(3)&(4)).
- 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: intuitively, we know can only make $R(D)$ smaller by making $D$ larger.
- For the moment, lets ignore the inequality constraint $q(\hat{x}|x) \geq 0$ and hope that we find everywhere positive solutions.
Computing $R(D)$

- Both inequality (Eq.(2)) and equality constraints (Eqs.(3)&(4)).
- 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: intuitively, we know can only make $R(D)$ smaller by making $D$ larger.
- For the moment, lets 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:

$$J(Q) = \sum_{x, \hat{x}} p(x)q(\hat{x}|x) \log \frac{q(\hat{x}|x)}{q(\hat{x})}$$

$$+ \lambda \left( \sum_{x, \hat{x}} p(x)q(\hat{x}|x)d(x, \hat{x}) - D \right)$$

$$+ \sum_{x} \nu(x) \left( \sum_{\hat{x}} q(\hat{x}|x) - 1 \right)$$

(8)
After some algebra, we get:

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

(9)
Computing $R(D)$

- After some algebra, we get:

$$ \frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x}) \mu(x)} + \lambda d(x, \hat{x}) \right] = 0 \quad (9) $$

- Implying that

$$ q(\hat{x}|x) = q(\hat{x}) e^{-\lambda d(x, \hat{x})} \sum \hat{y} q(\hat{y}) e^{-\lambda d(x, \hat{y})} \quad (10) $$
Computing $R(D)$

- After some algebra, we get:

$$\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] = 0$$  \hspace{1cm} (9)

- Implying that

$$q(\hat{x}|x) = q(\hat{x}) e^{-\lambda d(x, \hat{x})} \mu(x)$$  \hspace{1cm} (10)
Computing $R(D)$

- After some algebra, we get:

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

(9)

- Implying that

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

(10)
After some algebra, we get:

\[
\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] = 0
\]  \hspace{1cm} (9)

Implying that

\[
q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x, \hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x, \hat{y})}}
\]  \hspace{1cm} (10)
After some algebra, we get:

\[
\frac{\partial J}{\partial q(\hat{x}|x)} = p(x) \left[ \log \frac{q(\hat{x}|x)}{q(\hat{x})\mu(x)} + \lambda d(x, \hat{x}) \right] = 0 \tag{9}
\]

Implying that

\[
q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \tag{10}
\]

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.
The update is:

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

(11)
Computing $R(D)$

- The update is:

$$q(\hat{x}|x) = q(\hat{x}|x) e^{-\lambda d(x,\hat{x})} \mu(x)$$

Note that $\mu(x) = \sum \hat{y} q(\hat{y}|x) e^{-\lambda d(x,\hat{y})}$ 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 $R(D)$

The update is:

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

(11)
Computing $R(D)$

The update is:

$$q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum \hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})}}$$  \hspace{1cm} (11)
The update is:

\[
q(\hat{x} | x) = \frac{q(\hat{x}) e^{-\lambda d(x, \hat{x})}}{\mu(x)} = \frac{q(\hat{x}) e^{-\lambda d(x, \hat{x})}}{\sum_{\hat{y}} q(\hat{y}) e^{-\lambda d(x, \hat{y})}} \tag{11}
\]

Note that \(\mu(x) = \sum_{\hat{y}} q(\hat{y}) e^{-\lambda d(x, \hat{y})}\) since \(\sum_{\hat{x}} q(\hat{x} | x) = 1\).
Computing $R(D)$

The update is:

$$q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}}$$  \hspace{1cm} (11)

Note that $\mu(x) = \sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}$ since $\sum_{\hat{x}} q(\hat{x}|x) = 1$.

If $d(x, \hat{x})$ is large, then $q(\hat{x}|x)$ will be small.
Computing $R(D)$

- The update is:

$$q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}}$$  \hspace{1cm} (11)

- Note that $\mu(x) = \sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}$ 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.
The update is:

\[ q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\mu(x)} = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum\hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \]  (11)

Note that \( \mu(x) = \sum\hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})} \) 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 $q(\hat{x})$, we find $q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$, yielding:

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

(13)
Computing $R(D)$

To solve for $q(\hat{x})$, we find

$$q(\hat{x}) = \sum_x p(x)q(\hat{x}|x),$$

yielding:

$$q(\hat{x})$$

Equation (13)
To solve for $q(\hat{x})$, we find

$$q(\hat{x}) = \sum_x p(x) q(\hat{x} | x),$$

yielding:

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

(12)

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

(13)
Computing $R(D)$

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

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

(12)

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

(13)
To solve for \( q(\hat{x}) \), we find 

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

yielding:

\[
q(\hat{x}) = \sum_x p(x) \left( \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum \hat{y} q(\hat{y}) e^{-\lambda d(x,\hat{y})}} \right) \tag{12} 
\]

\[
= \frac{\sum_x p(x)e^{-\lambda d(x,\hat{x})}}{\sum \hat{y} q(\hat{y}) e^{-\lambda d(x,\hat{y})}} \tag{13} 
\]
To solve for \( q(\hat{x}) \), we find \( q(\hat{x}) = \sum_x p(x)q(\hat{x}|x) \), yielding:

\[
q(\hat{x}) = \sum_x p(x) \left( \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \right) \tag{12}
\]

\[
= q(\hat{x}) \frac{\sum_x p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \tag{13}
\]

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

\[
C(\hat{x}) = \sum_x \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} = 1 \tag{14}
\]
To solve for $q(\hat{x})$, we find $q(\hat{x}) = \sum_x p(x)q(\hat{x}|x)$, yielding:

$$q(\hat{x}) = \sum_x p(x) \left( \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum \hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \right)$$  \hspace{1cm} (12)

$$= \frac{q(\hat{x})}{\sum \hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \sum_x p(x)e^{-\lambda d(x,\hat{x})}$$  \hspace{1cm} (13)

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

$$C(\hat{x}) = \sum_x \frac{p(x)e^{-\lambda d(x,\hat{x})}}{\sum \hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})}} = 1$$  \hspace{1cm} (14)

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 $\lambda$. 

Prof. Jeff Bilmes
According to Eq. (11), as long as $q(\hat{x}) > 0$ then $q(\hat{x}|x) > 0$ as well.
According to Eq. (11), 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.
Computing $R(D)$

**Theorem 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.
More intuition: From previous definition, we have

\[ q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \quad (15) \]
More intuition: From previous definition, we have

\[ q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum \hat{y} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \]  \hspace{1cm} (15)

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.
More intuition: From previous definition, we have

\[ q(\hat{x}|x) = \frac{q(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} q(\hat{y})e^{-\lambda d(x,\hat{y})}} \]  \hspace{1cm} (15)

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 \( s = -\lambda \).
Computing $R(D)$

**Theorem 2.2**

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.

$$R' = \frac{dR}{dD} \bigg|_{D_s} = s \quad (16)$$

**Proof.**

Take derivatives and use the chain rule . . .
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.

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

(16)

**Proof.**

Take derivatives and use the chain rule . . .

- Pictorially,
Thus, we have a way to compute $R(D)$ in principle for any $s = -\lambda$. 
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.
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.
- We continue this process until all are positive.
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.
- We continue this process until all are positive.
- If we have only one left, then we have a $R = 0$ case.
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.
- 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).
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.
- 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 this.
Consider the problem: we have two convex sets $A, B \subseteq \mathcal{R}^n$. 

- Consider the following algorithm:
  1. Chose $a_0 \in A$ arbitrarily;
  2. for $n = 1, \ldots$ do
  3. Choose $b_n \in \text{argmin}_{b \in B} d(a_n - 1, b)$;
  4. Choose $a_n \in \text{argmin}_{a \in A} d(a, b_n)$;
2 convex sets

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

- 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:

$$d_{\text{min}} = \min_{a \in A, b \in B} d(a, b) \quad (17)$$
2 convex sets

- 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:

$$d_{\text{min}} = \min_{a \in A, b \in B} d(a, b) \quad (17)$$

- 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 $a_n \in \arg\min_{a \in A} d(a, b_n)$;
Theorem 2.3

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

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

\[
D(p(x)p(y|x)||p(x)r^*(y)) = \min_{r(y) \in \Delta} D(p(x)p(y|x)||p(x)r(y)) \tag{18}
\]

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

\[
\max_{r(x|y) \in \Delta^2} \sum_{x,y} p(x)p(y|x) \log \frac{r(x|y)}{p(x)} = \sum_{x,y} p(x)p(y|x) \log \frac{r^*(x|y)}{p(x)} \tag{19}
\]
This then gives $R(D)$ in the alternating minimization form:

$$R(D) = \min_{q \in B} \min_{p \in A} D(p || q)$$  \hspace{1cm} (20)

where

$$A = \{ q(x, \hat{x}) : q(x, \hat{x}) = p(x)r(\hat{x}) \text{ for arbitrary } r(\hat{x}) \}$$  \hspace{1cm} (21)

$$B = \left\{ p(x, \hat{x}) : p(x, \hat{x}) = q(\hat{x}|x)p(x) \text{ s.t. } \sum_{x,y} p(x, \hat{x})d(x, \hat{x}) \leq D \right\}$$  \hspace{1cm} (22)
So, to compute $R(D)$ at some point $s = -\lambda$, start with some arbitrary $r(\hat{x})$, and find the corresponding $q(\hat{x}|x)$. 

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

$$q(\hat{x}|x) = \frac{r(\hat{x}) e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} r(\hat{y}) e^{-\lambda d(x,\hat{y})}}$$  \hspace{1cm} (23)
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

$$q(\hat{x}|x) = \frac{r(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} r(\hat{y})e^{-\lambda d(x,\hat{y})}}$$ (23)

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

$$r(\hat{x}) = \sum_{x} p(x)q(\hat{x}|x)$$ (24)
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
  \[
  q(\hat{x}|x) = \frac{r(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} r(\hat{y})e^{-\lambda d(x,\hat{y})}} \tag{23}
  \]
- Once we have $q(\hat{x}) = q(\hat{x}|x)p(x)$, we find corresponding next $r(\hat{x})$ from the projection
  \[
  r(\hat{x}) = \sum_{x} p(x)q(\hat{x}|x) \tag{24}
  \]
- We repeat this alternating projection/minimization procedure until convergence.
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

$$q(\hat{x}|x) = \frac{r(\hat{x})e^{-\lambda d(x,\hat{x})}}{\sum_{\hat{y}} r(\hat{y})e^{-\lambda d(x,\hat{y})}} \quad (23)$$

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

$$r(\hat{x}) = \sum_{x} p(x)q(\hat{x}|x) \quad (24)$$

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

- In this case, we have:

\[
C = \max_{q(x|y)} \max_{r(x)} \sum_{x,y} r(x) p(y|x) \log \frac{q(x|y)}{r(y)}
\]  \hspace{1cm} (25)

- We guess a 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} (26)
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 $x \in \mathcal{X}$, $p(x) \geq 0$. 

---

Prof. Jeff Bilmes  
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 $P_i \in \mathcal{P}$ arbitrarily.

Then the result we will get is that:

$D(\mathcal{P}_n || \mathcal{Q}_n) \to \inf_{(P, Q) \in (\mathcal{P}_0, \mathcal{Q})} D(\mathcal{P} || \mathcal{Q})$ (29)

$P_n \to P^*$, $Q_n \to Q^*$ sometimes as well.

$\mathcal{P}_0$ are the entries of $\mathcal{P}$ that we care about.
Summary

- 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 $P_i \in \mathcal{P}$ arbitrarily.
- Define $Q_n \in \text{argmin}_{Q \in \mathcal{Q}} D(P_n || Q)$.
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 $P_i \in \mathcal{P}$ arbitrarily.
- Define $Q_n \in \arg\min_{Q \in \mathcal{Q}} D(P_n \| Q)$.

That is, we have the following procedure:

$$Q_n \in \arg\min_{Q \in \mathcal{Q}} D(P_n \| Q)$$  \hspace{1cm} (27)

$$P_{n+1} \in \arg\min_{P \in \mathcal{P}} D(P \| Q_n)$$  \hspace{1cm} (28)
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 \( P_i \in \mathcal{P} \) arbitrarily.
- Define \( Q_n \in \arg\min_{Q \in \mathcal{Q}} D(P_n||Q) \).
- That is, we have the following procedure:

\[
Q_n \in \arg\min_{Q \in \mathcal{Q}} D(P_n||Q) \\
P_{n+1} \in \arg\min_{P \in \mathcal{P}} D(P||Q_n)
\]  

(27)  

(28)

- Then the result we will get is that:

\[
D(P_n||Q_n) \rightarrow \inf_{(P,Q) \in (\mathcal{P}_0, \mathcal{Q})} D(P||Q)
\]  

(29)

where \( \mathcal{P}_0 = \{ P \in \mathcal{P} : D(P||Q_n) < \infty \text{ for some } n \} \) and \( P_n \rightarrow P^*, \)
\( Q_n \rightarrow Q^* \) sometimes as well.
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 $P_i \in \mathcal{P}$ arbitrarily.

Define $Q_n \in \argmin_{Q \in \mathcal{Q}} D(P_n \| Q)$.

That is, we have the following procedure:

$$Q_n \in \argmin_{Q \in \mathcal{Q}} D(P_n \| Q)$$  \hspace{1cm} (27)

$$P_{n+1} \in \argmin_{P \in \mathcal{P}} D(P \| Q_n)$$  \hspace{1cm} (28)

Then the result we will get is that:

$$D(P_n \| Q_n) \rightarrow \inf_{(P,Q) \in (\mathcal{P}_0, \mathcal{Q})} D(P \| Q)$$  \hspace{1cm} (29)

where $\mathcal{P}_0 = \{ P \in \mathcal{P} : D(P \| Q_n) < \infty \text{ for some } n \}$ and $P_n \rightarrow P^*$, $Q_n \rightarrow Q^*$ 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.
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.).
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).
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)
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.
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.
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.).
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 $d : \mathcal{P} \times \mathcal{Q} \rightarrow \mathbb{R} \cup \{+\infty\}$ which need not be KL-divergence.
Properties of $d$

- Let $d(P, Q)$ be an 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 $\infty$).
Properties of $d$

- Let $d(P, Q)$ be an 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 $P \overset{1}{\rightarrow} Q'$ 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 $Q'$. 
Properties of $d$

- Let $d(P, Q)$ be an 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 $P \overset{1}{\rightarrow} Q'$ 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 $Q \overset{2}{\rightarrow} P'$, indicating we minimize over $P$, holding the 2nd argument $Q$ fixed.
Properties of $d$

- Let $d(P, Q)$ be an 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 $P \overset{1}{\rightarrow} Q'$ 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 $Q \overset{2}{\rightarrow} P'$, indicating we minimize over $P$, holding the 2nd argument $Q$ fixed.

- Sequences obtained by alternating minimization $\{(P_n, Q_n)\}_{n=0}^{\infty}$ as:

$$
P_0 \overset{1}{\rightarrow} Q_0 \overset{2}{\rightarrow} P_1 \overset{1}{\rightarrow} Q_1 \overset{2}{\rightarrow} P_2 \overset{1}{\rightarrow} Q_2 \overset{2}{\rightarrow} P_3 \overset{1}{\rightarrow} Q_3 \overset{2}{\rightarrow} \cdots \tag{30}
$$

where we start arbitrarily with $P_0$. 

Goal: sufficient conditions for the convergence of the alternating minimization procedure.
Properties of $d$

- Let $d(P, Q)$ be an 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 $P \ x{1} \rightarrow Q'$ 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 $Q \ x{2} \rightarrow P'$, indicating we minimize over $P$, holding the 2nd argument $Q$ fixed.

- Sequences obtained by alternating minimization $\{(P_n, Q_n)\}_{n=0}^\infty$ as:

  $$P_0 \ x{1} \rightarrow Q_0 \ x{2} \rightarrow P_1 \ x{1} \rightarrow Q_1 \ x{2} \rightarrow P_2 \ x{1} \rightarrow Q_2 \ x{2} \rightarrow P_3 \ x{1} \rightarrow Q_3 \ x{2} \rightarrow \cdots$$  

  (30)

where we start arbitrarily with $P_0$.

- Goal: sufficient conditions for the convergence of the alternating minimization procedure.
Definition 3.1 (Five Points Property (5PP))

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

$$d(P, Q) + d(P, Q_0) \geq d(P, Q_1) + d(P_1, Q_1)$$  \hspace{1cm} (31)

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 $\mathcal{Q}$.
Five Points Property

Definition 3.1 (Five Points Property (5PP))

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

$$
    d(P, Q) + d(P, Q_0) \geq d(P, Q_1) + d(P_1, Q_1) \quad (31)
$$

whenever $Q_0 \rightarrow P_1 \rightarrow Q_1$. $d(\cdot, \cdot)$ 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 😊).
Five Points Property

**Definition 3.1 (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:

$$ \forall Q \in \mathcal{Q}, \forall Q_0 \in \mathcal{Q}, \text{ we have:} $$

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

(31)

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 😊).
- 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 $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$
Five Points Property

\[ P \in \mathcal{P} \]
\[ \forall Q \in \mathcal{Q}, Q_0 \in \mathcal{Q} \]
\[
\begin{align*}
    d(P, Q) + d(P, Q_0) & \geq d(P, Q_1) + d(P_1, Q_1) \\

    P_1 & \in \arg\min_{P \in \mathcal{P}} d(P, Q_0) \\
    Q_1 & \in \arg\min_{Q \in \mathcal{Q}} d(P_1, Q)
\end{align*}
\]
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

$$
\lim_{n \to \infty} d(P_n, Q_n) = \inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} d(P, Q) = d_{\min} \tag{32}
$$

as long as

$$
d_{\min} = \inf_{P \in \mathcal{P}_0, Q \in \mathcal{Q}} d(P, Q) \tag{33}
$$

where

$$
\mathcal{P}_0 = \{ P : P \in \mathcal{P}, d(P, Q_n) < \infty \text{ for some } n \} \tag{34}
$$
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

\[
\lim_{n \to \infty} d(P_n, Q_n) = \inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} d(P, Q) = d_{\text{min}} \tag{32}
\]

as long as

\[
d_{\text{min}} = \inf_{P \in \mathcal{P}_0, Q \in \mathcal{Q}} d(P, Q) \tag{33}
\]

where

\[
\mathcal{P}_0 = \{ P : P \in \mathcal{P}, d(P, Q_n) < \infty \text{ for some } n \} \tag{34}
\]

- 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 $B \subseteq \mathcal{Q}$,

\[ d(A, B) \triangleq \inf_{P \in A, Q \in B} d(P, Q) \quad (35) \]

Since $d(P, Q) \in \mathbb{R} \cup \{+\infty\}$, $d(A, B)$ does not take the value $-\infty$. 
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} (35)$$

Since $d(P, Q) \in \mathbb{R} \cup \{+\infty\}$, $d(A, B)$ does not take the value $-\infty$. 
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{2cm} (35)

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

Lemma 3.2

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

$$d(P_n, Q_n) \geq d(P_0, Q) \ \forall n$$  \hspace{2cm} (36)
Definitions

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

$$d(A, B) \triangleq \inf_{P \in A, Q \in B} d(P, Q) \quad (35)$$

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

Lemma 3.2

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

$$d(P_n, Q_n) \geq d(\mathcal{P}_0, Q) \quad \forall n \quad (36)$$

Proof.

Obvious via definitions.
Definitions

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

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

(35)

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

Lemma 3.2

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

\[
   d(P_n, Q_n) \geq d(P_0, Q) \quad \forall n
\]  

(36)

Proof.

Obvious via definitions.

Our goal is to first find when \( \lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \).
Recall limsup/liminf

- Recall,

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

(37)

where

\[ S = \{ a : a = \sup B_n \text{ for some } n, \text{ with } B_n = \{ a_n, a_{n+1}, \ldots, \} \}. \]
Recall limsup/liminf

- Recall,

\[ \limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S \tag{37} \]

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, \( \limsup_{x \to \infty} \sin(x) = 1 \).
recall \( \limsup / \liminf \)

- Recall,

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

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, \( \limsup_{x \to \infty} (\sin(x) - \sin^2(x)) = \)

\[
\text{(37)}
\]
Recall limsup/liminf

Recall,

$$\limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S$$ \hspace{1cm} (37)

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, $$\limsup_{x \to \infty} (\sin(x) - \sin^2(x)) =$$
Recall, 

\[ \limsup_{n \to \infty} a_n \triangleq \inf_{n>0} \left( \sup_{k>n} a_k \right) = \inf S \tag{37} \]

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).
Recall limsup/liminf

Recall,

\[
\limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S \tag{37}
\]

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,

\[
\liminf_{n \to \infty} a_n \triangleq \sup_{n > 0} \left( \inf_{k > n} a_k \right) \tag{38}
\]

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

Lemma 3.3

Let $a_n, b_n$ for $n = 0, 1, \ldots$ be extended real sequences in the sense
$\forall n, a_n, b_n \in \mathbb{R} \cup \{+\infty\}$. 

Key Lemma

Lemma 3.3

Let $a_n, b_n$ for $n = 0, 1, \ldots$ be extended real sequences in the sense $orall n, a_n, b_n \in \mathbb{R} \cup \{+\infty\}$. Let $c$ be finite arbitrary such that:

$$c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots.$$ (39)
Key Lemma

Lemma 3.3

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:

$$c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots.$$  \hspace{1cm} (39)

And also assume that

$$\limsup_{n \to \infty} b_n > -\infty, \quad \text{and} \quad \exists n_0 \text{ s.t. } b_{n_0} < \infty.$$ \hspace{1cm} (40)

Then

$$\liminf_{n \to \infty} a_n \leq c.$$  \hspace{1cm} (41)

Also, if in addition, we assume that

$$\sum_{n=0}^{\infty} (c - a_n) + < \infty$$

then

$$\sum_{n=0}^{\infty} |a_n - c| < \infty$$  \hspace{1cm} (42)

and as a result

$$\lim_{n \to \infty} a_n = c.$$  \hspace{1cm} (43)
Key Lemma

Lemma 3.3

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:
\[
c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots.
\] (39)

And also assume that
\[
\limsup_{n \to \infty} b_n > -\infty, \quad \text{and} \quad \exists n_0 \text{ s.t. } b_{n_0} < \infty.
\] (40)

Then
\[
\liminf_{n \to \infty} a_n \leq c
\] (41)

And if in addition we assume that
\[
\infty \sum_{n=0}^{\infty} (c - a_n) + < \infty
\]
then
\[
\infty \sum_{n=0}^{n_0+1} |a_n - c| < \infty
\] (42)

and as a result
\[
\lim_{n \to \infty} a_n = c
\] (43)
Key Lemma

Lemma 3.3

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:

$$c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots.$$  \hfill (39)

And also assume that

$$\limsup_{n \to \infty} b_n > -\infty, \quad \text{and} \quad \exists n_0 \text{ s.t. } b_{n_0} < \infty.$$  \hfill (40)

Then

$$\liminf_{n \to \infty} a_n \leq c$$  \hfill (41)

Also, if in addition, we assume that

$$\sum_{n=0}^{\infty} (c - a_n)^+ < \infty$$  \hfill (42)
Key Lemma

Lemma 3.3

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:
\[ c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots. \] (39)

And also assume that
\[ \limsup_{n \to \infty} b_n > -\infty, \quad \text{and} \quad \exists n_0 \text{ s.t. } b_{n_0} < \infty. \] (40)

Then
\[ \liminf_{n \to \infty} a_n \leq c \] (41)

Also, if in addition, we assume that
\[ \sum_{n=0}^{\infty} (c - a_n)^+ < \infty \quad \text{then} \quad \sum_{n=n_0+1}^{\infty} |a_n - c| < \infty \] (42)
**Key Lemma**

**Lemma 3.3**

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:

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

(39)

And also assume that

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

(40)

Then

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

(41)

Also, if in addition, we assume that

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

(42)

and as a result

\[ \lim_{n \to \infty} a_n = c \]  

(43)
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 $\lim \inf_{n \to \infty} a_n > c$ since if so, eventually we'd get $(c - a_n) = \infty$ and the sum would be finite).

Thus, $\lim \inf_{n \to \infty} a_n \leq c$. 

...
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 $a_n < \infty$. 

...
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 \( \lim \inf_{n \to \infty} a_n > c \) since if so, eventually we'd get \( (c - a_n)^+ \) and the sum would be finite).
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 \( \lim \inf_{n \to \infty} a_n > c \) since if so, eventually we’d get \( (c - a_n)^+ \) and the sum would be finite).

Thus, \( \lim \inf_{n \to \infty} a_n \leq c \).
Proof.

- Next if \( b_{n_0} < \infty \) for some \( n_0 \), then since \( c \) is finite, and since

\[
c + b_{n-1} \geq b_n + a_n,
\]

then we have \( a_n < \infty, b_n < \infty, \forall n > n_0 \).
Proof.

- Next if $b_{n_0} < \infty$ for some $n_0$, then since $c$ is finite, and since

$$c + b_{n-1} \geq b_n + a_n,$$  \hspace{1cm} (44)

then we have $a_n < \infty$, $b_n < \infty$, $\forall n > n_0$.

- Thus, $a_n - c \leq b_{n-1} - b_n$ for $n > n_0$,
Proof.

Next if $b_{n_0} < \infty$ for some $n_0$, then since $c$ is finite, and since

$$c + b_{n-1} \geq b_n + a_n,$$  \hspace{1cm} (44)

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:

$$\sum_{n=n_0+1}^{n} (a_n - c) \leq \sum_{n=n_0+1}^{n} (b_{n-1} - b_n) = b_{n_0} - b_n \ \forall n > n_0 \ \hspace{1cm} (45)$$
Next if $b_{n_0} < \infty$ for some $n_0$, then since $c$ is finite, and since
\[ c + b_{n-1} \geq b_n + a_n, \] (44)
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:
\[ \sum_{n=n_0+1}^{n} (a_n - c) \leq \sum_{n=n_0+1}^{n} (b_{n-1} - b_n) = b_{n_0} - b_n \quad \forall n > n_0 \] (45)

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 $\lim_{n\to\infty} b_n - b_{n_0} > -\infty$, or $\lim_{n\to\infty} b_{n_0} - b_n < \infty$, meaning that it has a limit and $\sum_{n=n_0+1}^{\infty} (c - a_n) < \infty$. 

...
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 \( \sum_{n=n_0+1}^{\infty} |a_n - c| < \infty \).
### 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 $|a| = a^+ + a^- = -c_\pm < \infty$. 

...
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 \( \lim_{n\to\infty} a_n = c \).
**Proof.**

- 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.)

  (*exercise: justify this step*)
Proof.

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. (exercise: justify this step)

Also, if the limit is finite, then we have

$$\sum_{n=n_0+1}^{\infty} (a_n - c) < \infty \Rightarrow \sum_{n=n_0+1}^{\infty} (a_n - c)^+ < \infty \Rightarrow \sum_{n=n_0+1}^{\infty} (c - a_n)^- < \infty$$
Key Lemma

Proof.

- 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). (exercise: justify this step)

- 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

$$\sum_{n=n_0+1}^{\infty}(c - a_n)^- + (c - a_n)^+ < \infty$$
Key Lemma

Proof.

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). (exercise: justify this step)

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 \( \lim_{n \to \infty} a_n = c \).
Theorem 3.4

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:

\[
d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n) \quad n = 1, 2, \ldots \quad (46)
\]
Theorem 3.4

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:

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

(46)

Note: no minimization done here, only 5PP condition on the sequences.
Theorem 3.4

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

\[
d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n) \quad n = 1, 2, \ldots \tag{46}
\]

Note: no minimization done here, only 5PP condition on the sequences. Then if either: A) \( \forall P \in \mathcal{P}_0 \);
1st Main theorem

Theorem 3.4

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

\[
d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n) \quad n = 1, 2, \ldots \tag{46}
\]

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. \( d(P, Q) = d(\mathcal{P}_0, Q) \),
1st Main theorem

Theorem 3.4

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:

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

(46)

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:

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

(47)
1st Main theorem

**Theorem 3.4**

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:

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

(46)

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:

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

(47)

And if A holds then \( d(P_n, Q_n) \) is non-increasing.
Theorem 3.4

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:

\[
d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n) \quad n = 1, 2, \ldots
\] (46)

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:

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q).
\] (47)

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

\[
\sum_{n=n_1}^{\infty} (d(P_n, Q_n) - d(P_0, Q)) < \infty
\] (48)
Sequences

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

\[
P_0 \to Q_0 \to P_1 \to Q_1 \to P_2 \to Q_2 \to P_3 \to Q_3 \to \cdots \tag{49}
\]
Sequences

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

\[
P_0 \rightarrow Q_0 \rightarrow P_1 \rightarrow Q_1 \rightarrow P_2 \rightarrow Q_2 \rightarrow P_3 \rightarrow Q_3 \rightarrow \cdots \quad (49)
\]

- Then we have that:

\[
d(P_n, Q_n) \geq d(P_{n+1}, Q_n) \geq d(P_{n+1}, Q_{n+1}) \quad \text{for } n = 0, 1, \ldots \quad (50)
\]
Sequences

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

\[
P_0 \xrightarrow{1} Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \xrightarrow{2} P_2 \xrightarrow{1} Q_2 \xrightarrow{2} P_3 \xrightarrow{1} Q_3 \xrightarrow{2} \cdots \tag{49}
\]

- Then we have that:

\[
d(P_n, Q_n) \geq d(P_{n+1}, Q_n) \geq d(P_{n+1}, Q_{n+1}) \quad \text{for } n = 0, 1, \ldots \tag{50}
\]

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

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

\[
P_0 \rightarrow Q_0 \rightarrow P_1 \rightarrow Q_1 \rightarrow P_2 \rightarrow Q_2 \rightarrow P_3 \rightarrow Q_3 \rightarrow \cdots
\]

(49)

- Then we have that:

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

(50)

- 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 3.4 met at \( P \in \mathcal{P}_0 \)
Sequences

• 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.
Sequences

- That is, for $n = 1$ we have $Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1$ so
  
  \[ d(P, Q) + d(P, Q_0) \geq d(P, Q_1) + d(P_1, Q_1) \quad \forall Q, Q_0 \]  
  (51)

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

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

\[ \text{Prof. Jeff Bilmes} \]

Sequences

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

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

$$d(P, Q) + d(P, Q_0) \geq d(P, Q_1) + d(P_1, Q_1) \quad \forall Q, Q_0$$  \hspace{1cm} (51)

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

- For $n = 2$ we have $Q_1 \xrightarrow{2} P_2 \xrightarrow{1} Q_2$ so

$$d(P, Q) + d(P, Q_1) \geq d(P, Q_2) + d(P_2, Q_2) \quad \forall Q, Q_0$$  \hspace{1cm} (52)

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

Sequences

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

\[
d(P, Q) + d(P, Q_0) \geq d(P, Q_1) + d(P_1, Q_1) \quad \forall Q, Q_0
\] (51)

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

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

\[
d(P, Q) + d(P, Q_1) \geq d(P, Q_2) + d(P_2, Q_2) \quad \forall Q, Q_0
\] (52)

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

- Same for \( n > 2 \), etc.
So Theorem 3.4 holds in this case (i.e., \( \lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \)).
Sequences

- So Theorem 3.4 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.
Sequences

- So Theorem 3.4 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.
Sequences

- So Theorem 3.4 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.
Sequences

- So Theorem 3.4 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.
Definition 3.5 (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 $d : \mathcal{P} \times \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ and $\delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}_+$, the three points property for $P \in \mathcal{P}$ holds if $\forall Q_0$

$$
\delta(P, P_1) + d(P_1, Q_0) \leq d(P, Q_0) \quad \text{whenever } Q_0 \xrightarrow{2} P_1
$$

(53)
Three Points Property

Definition 3.5 (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 $\forall Q_0$

$$\delta(P, P_1) + d(P_1, Q_0) \leq d(P, Q_0) \text{ whenever } Q_0 \xrightarrow{2} P_1$$

(53)

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

\[ P \in \mathcal{P} \]
\[ \forall Q_0 \in \mathcal{Q} \]
\[ P_1 \in \arg\min_{P \in \mathcal{P}} d(P, Q_0) \]

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

\[ P_1 \in \arg\min_{P \in \mathcal{P}} d(P, Q_0) \]
Four Points Property (4PP)

Definition 3.6 (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} (54)
Four Points Property (4PP)

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

\[
P \in \mathcal{P} \\
\forall Q \in \mathcal{Q}, \forall P_1 \in \mathcal{P}
\]

\[
Q_1 \in \arg\min_{Q \in \mathcal{Q}} d(P_1, Q)
\]
Second Main Theorem

Theorem 3.7

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

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

(55)

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,
Theorem 3.7

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

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \quad (55)
\]

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

\[ 3PP + 4PP \Rightarrow 5PP \]
Theorem 3.7

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

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \tag{55}
\]

if \( P \) is defined by either: A) all \( P \in P_0 \); or B) some \( P \in P_0 \) 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 3.4.
Second Main Theorem

Proof.

- We saw that 5PP + alternating minimization implies Theorem 3.4.
- Combining 3PP and 4PP we have:

\[
Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \tag{56}
\]

\[
d(P, Q_0) - \delta(P, P_1) \geq d(P_1, Q_0) \quad \text{3PP} \tag{57}
\]

\[
\delta(P, P_1) + d(P, Q) \geq d(P, Q_1) \quad \text{4PP} \tag{58}
\]
Second Main Theorem

Proof.

- We saw that 5PP + alternating minimization implies Theorem 3.4.
- Combining 3PP and 4PP we have:

\[ Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \]  
\[ d(P, Q_0) - \delta(P, P_1) \geq d(P_1, Q_0) \]  \hspace{1cm} (56) \hspace{1cm} 3PP
\[ \delta(P, P_1) + d(P, Q) \geq d(P, Q_1) \]  \hspace{1cm} (57) \hspace{1cm} 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 \( Q_0 \xrightarrow{2} P_1 \).
Second Main Theorem

Proof.

- We saw that 5PP + alternating minimization implies Theorem 3.4.
- Combining 3PP and 4PP we have:

\[ Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \]  

(56)

\[ d(P, Q_0) - \delta(P, P_1) \geq d(P_1, Q_0) \quad \text{3PP} \]  

(57)

\[ \delta(P, P_1) + d(P, Q) \geq d(P, Q_1) \quad \text{4PP} \]  

(58)

- 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:

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

(59)

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

(60)

since \( P_1 \xrightarrow{1} Q_1 \), thus giving 5PP.
Second Main Theorem

Proof.

Further, if 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

\[ \delta(P, P_{n+1}) + d(P_{n+1}, Q_n) \leq d(P, Q_n) \leq \delta(P, P_n) + d(P, Q) \] (61)
Second Main Theorem

Proof.

- Further, if 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

\[ \delta(P, P_{n+1}) + d(P_{n+1}, Q_n) \leq d(P, Q_n) \leq \delta(P, P_n) + d(P, Q) \quad (61) \]

- This implies

\[ \delta(P, P_{n+1}) \leq \delta(P, P_n) + [d(P, Q) - d(P_{n+1}, Q_n)] \forall Q \quad (62) \]

so

\[ \delta(P, P_{n+1}) \leq \delta(P, P_n) + \underbrace{[d(P, Q) - d(P_{n+1}, Q_n)]}_{\leq 0} \quad (63) \]
Proof.

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

Proof.

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

- Note, this shows that:

$$\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \quad (64)$$
Second Main Theorem

Proof.

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

- Note, this shows that:

$$\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q)$$ (64)

- Ideally, we would like $d(P_0, Q) = d(P, Q)$
Second Main Theorem

Proof.

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

- Note, this shows that:

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \tag{64}
\]

- 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.
Second Main Theorem

Proof.

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

Note, this shows that:

$$\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q)$$  \hspace{1cm} (64)

- 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

$$\mathcal{P}_0 = \{P : D(P||Q) < +\infty\}$$

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

Let $P, Q$ be closed convex subsets of a Hilbert space (normed space with a dot product s.t., every Cauchy sequence converges). Assume, e.g., $\mathbb{R}^n$
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 $\delta(P, P') = \| P - P' \|^2$. 
Let $P, Q$ 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

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

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

- Thus, it satisfies 5PP.