EE515A – Information Theory II
Spring 2012

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
Spring Quarter, 2012
http://j.ee.washington.edu/~bilmes/classes/ee515a_spring_2012/

Lecture 28 - May 2nd, 2012
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.
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.
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 \( x \in \mathcal{X}, p(x) \geq 0 \)
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 $P_0 \in \mathcal{P}$ arbitrarily.
Summary

- Let $\mathcal{P}, Q$ 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 $P_0 \in \mathcal{P}$ arbitrarily.
- Define $Q_n \in \text{argmin}_{Q \in Q} D(P_n || Q)$.
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 $P_0 \in \mathcal{P}$ arbitrarily.
- Define $Q_n \in \arg\min_{Q \in \mathcal{Q}} D(P_n \| Q)$.
- That is, we have the following procedure:

\[
Q_n \in \arg\min_{Q \in \mathcal{Q}} D(P_n \| Q) \quad (1)
\]
\[
P_{n+1} \in \arg\min_{P \in \mathcal{P}} D(P \| Q_n) \quad (2)
\]
Summary

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 $P_0 \in \mathcal{P}$ arbitrarily.

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

That is, we have the following procedure:

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

$$P_{n+1} \in \arg\min_{P \in \mathcal{P}} D(P \| Q_n)$$ (2)

Then the result we will get is that:

$$D(P_n \| Q_n) \to \inf_{(P, Q) \in (\mathcal{P}_0, \mathcal{Q})} D(P \| Q)$$ (3)

where $\mathcal{P}_0 = \{ P \in \mathcal{P} : D(P \| Q_n) < \infty \text{ for some } n \}$ and $P_n \to P^*$, $Q_n \to Q^*$ sometimes as well.
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 $P_0 \in \mathcal{P}$ 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} (1)

\[ P_{n+1} \in \arg\min_{P \in \mathcal{P}} D(P || Q_n) \]  \hspace{1cm} (2)

- Then the result we will get is that:

\[ D(P_n || Q_n) \to \inf_{(P,Q) \in (\mathcal{P}_0, \mathcal{Q})} D(P || Q) \]  \hspace{1cm} (3)

where $\mathcal{P}_0 = \{ P \in \mathcal{P} : D(P || Q_n) < \infty \text{ for some } n \}$ and $P_n \to P^*$, $Q_n \to Q^*$ sometimes as well.
- $\mathcal{P}_0$ are the entries of $\mathcal{P}$ that we care about.
Summary

This process has a geometric flavor, since it corresponds to alternating “projections” based on treating KL as a “distance” in some sense.
Summary

- This process has a geometric flavor, since it corresponds to alternating “projections” based on treating KL as a “distance” in some 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 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 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 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 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.
This process has a geometric flavor, since it corresponds to alternating “projections” based on treating KL as a “distance” in some 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 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 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 $\infty$).
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 \( Q' \).
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.
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:

$$P_0 \xrightarrow{1} Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \xrightarrow{2} P_2 \xrightarrow{1} Q_2 \xrightarrow{2} P_3 \xrightarrow{1} Q_3 \xrightarrow{2} \cdots \quad (4)$$

where we start arbitrarily with $P_0$. 

Properties of $d$

- Let $d(P, Q)$ be a half extended-real valued function. That is, for $P \in \mathcal{P}$, $Q \in \mathcal{Q}$, we have $d(P, Q) > -\infty$ (we exclude $-\infty$ but allow $\infty$).

- Also, $d(P, Q') = \min_{Q \in \mathcal{Q}} d(P, Q) < \infty$. This minimization is denoted as $P \overset{1}{\rightarrow} Q'$ 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:

$$P_0 \overset{1}{\rightarrow} Q_0 \overset{2}{\rightarrow} P_1 \overset{1}{\rightarrow} Q_1 \overset{2}{\rightarrow} P_2 \overset{1}{\rightarrow} Q_2 \overset{2}{\rightarrow} P_3 \overset{1}{\rightarrow} Q_3 \overset{2}{\rightarrow} \cdots \quad (4)$$

where we start arbitrarily with $P_0$.

- Goal: sufficient conditions for the convergence of the alternating minimization procedure.
Five Points Property

**Definition 2.1 (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:

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

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 \( \mathcal{Q} \).
Definition 2.1 (Five Points Property (5PP))

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

$$d(P, Q) + d(P, Q_0) \geq d(P, Q_1) + d(P_1, Q_1)$$ (5)

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 $Q$.
- It is a definition on sets of 5 points! (obviously 😊).
**Definition 2.1 (Five Points Property (5PP))**

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

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

whenever $Q_0 \overset{2}{\rightarrow} P_1 \overset{1}{\rightarrow} Q_1$. $d(\cdot, \cdot)$ 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 😊).
- **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

\[ P \in \mathcal{P} \]

\[ \forall Q \in \mathcal{Q}, Q_0 \in \mathcal{Q} \]

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

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

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

\[
\lim_{n \to \infty} d(P_n, Q_n) = \inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} d(P, Q) = d_{\text{min}} \tag{6}
\]

as long as

\[
d_{\text{min}} = \inf_{P \in \mathcal{P}_0, Q \in \mathcal{Q}} d(P, Q) \tag{7}
\]

where

\[
\mathcal{P}_0 = \{ P : P \in \mathcal{P}, d(P, Q_n) < \infty \text{ for some } n \} \tag{8}
\]
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

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

(6)

as long as

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

(7)

where

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

(8)

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

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

(9)

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

### Lemma 2.2

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

$$d(P_n, Q_n) \geq d(P_0, Q_0) \quad \forall n$$

(10)

**Proof.** Obvious via definitions.

Our goal is to first find when

$$\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q_0).$$
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} (9)

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 $B \subseteq \mathcal{Q}$,

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

(9)

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

**Lemma 2.2**

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

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

(10)
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} (9)

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

Lemma 2.2

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

$$d(P_n, Q_n) \geq d(\mathcal{P}_0, Q) \ \forall n$$  \hspace{1cm} (10)

Proof.

Obvious via definitions.

---

Prof. Jeff Bilmes

Definitions

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

\[
    d(A, B) \triangleq \inf_{P \in A, Q \in B} d(P, Q) \tag{9}
\]

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

Lemma 2.2

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

\[
    d(P_n, Q_n) \geq d(P_0, Q) \quad \forall n \tag{10}
\]

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,

\[
\limsup_{n \to \infty} a_n \triangleq \inf_{n>0} \left( \sup_{k>n} a_k \right) = \inf S \tag{11}
\]

where

\( S = \{ a : a = \sup B_n \text{ for some } n, \text{ with } B_n = \{ a_n, a_{n+1}, \ldots \} \} \).
Recall $\limsup$ and $\liminf$:

- Recall,

$$\limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S \quad (11)$$

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,

\[ \limsup_{n \to \infty} a_n \triangleq \inf_{n>0} \left( \sup_{k>n} a_k \right) = \inf S \quad (11) \]

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, \[ \limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S \] (11)

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/liminf

- Recall,

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

(11)

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, 

\[ \limsup_{n \to \infty} a_n \triangleq \inf_{n > 0} \left( \sup_{k > n} a_k \right) = \inf S \quad (11) \]

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,

\[ \liminf_{n \to \infty} a_n \triangleq \sup_{n > 0} \left( \inf_{k > n} a_k \right) \quad (12) \]

so \( \liminf \) asks for supremum convergence in the local minima.
Lemma 2.3

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 2.3

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:

\[
c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots.
\] (13)
Key Lemma

Lemma 2.3

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:

$$c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots.$$  \hspace{1cm} (13)

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} (14)
Key Lemma

Lemma 2.3

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:

\[
c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots \tag{13}
\]

And also assume that

\[
\limsup_{n \to \infty} b_n > -\infty, \quad \text{and} \quad \exists n_0 \text{ s.t. } b_{n_0} < \infty. \tag{14}
\]

Then

\[
\liminf_{n \to \infty} a_n \leq c \tag{15}
\]
Key Lemma

Lemma 2.3

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:
\[c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots.\] (13)

And also assume that
\[\limsup_{n \to \infty} b_n > -\infty, \quad \text{and} \quad \exists n_0 \text{ s.t. } b_{n_0} < \infty.\] (14)

Then
\[\liminf_{n \to \infty} a_n \leq c\] (15)

Also, if in addition, we assume that
\[\sum_{n=0}^{\infty} (c - a_n)^+ < \infty\] (16)
**Key Lemma**

**Lemma 2.3**

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:

\[
    c + b_{n-1} \geq b_n + a_n, \quad \text{for } n = 1, 2, \ldots. \tag{13}
\]

And also assume that

\[
    \limsup_{n \to \infty} b_n > -\infty, \quad \text{and} \quad \exists n_0 \text{ s.t. } b_{n_0} < \infty. \tag{14}
\]

Then

\[
    \liminf_{n \to \infty} a_n \leq c \tag{15}
\]

Also, if in addition, we assume that

\[
    \sum_{n=0}^{\infty} (c - a_n)^+ < \infty \quad \text{then} \quad \sum_{n=n_0+1}^{\infty} |a_n - c| < \infty \tag{16}
\]
Lemma 2.3

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:

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

(13)

And also assume that

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

(14)

Then

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

(15)

Also, if in addition, we assume that

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

(16)

and as a result

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

(17)
Theorem 2.4

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:

$$d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n) \quad n = 1, 2, \ldots$$  \hspace{1cm} (18)
1st Main theorem

Theorem 2.4

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:

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

(18)

Note: no minimization done here, only 5PP condition on the sequences.
Theorem 2.4

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:

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

Note: no minimization done here, only 5PP condition on the sequences. Then if either: A) \( \forall P \in P_0; \)

...
Theorem 2.4

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:

\[
d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n) \quad n = 1, 2, \ldots
\] (18)

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 2.4

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

\[
d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n) \quad n = 1, 2, \ldots \tag{18}
\]

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

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q). \tag{19}
\]
1st Main theorem

Theorem 2.4

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:

\[
d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n) \quad n = 1, 2, \ldots \tag{18}
\]

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:

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(\mathcal{P}_0, Q). \tag{19}
\]

And if A holds then \( d(P_n, Q_n) \) is non-increasing.
Theorem 2.4

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:

\[
d(P, Q) + d(P, Q_{n-1}) \geq d(P, Q_n) + d(P_n, Q_n) \quad n = 1, 2, \ldots \tag{18}
\]

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:

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(\mathcal{P}_0, Q). \tag{19}
\]

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

\[
\sum_{n=n_1}^{\infty} (d(P_n, Q_n) - d(\mathcal{P}_0, Q)) < \infty \tag{20}
\]
Proof.

- If $\mathcal{P}_0 = \emptyset$ then, for all $n \geq 1$, we have

$$d(P_n, Q_n) = d(\mathcal{P}_0, Q) = \inf_{P \in \mathcal{P}_0, Q \in Q} d(P, Q) = \inf \emptyset = \infty \quad (21)$$

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

- If $\mathcal{P}_0 = \emptyset$ then, for all $n \geq 1$, we have

$$d(P_n, Q_n) = d(\mathcal{P}_0, Q) = \inf_{P \in \mathcal{P}_0, Q \in \mathcal{Q}} d(P, Q) = \inf \emptyset = \infty \quad (21)$$

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 (18) holds for some $P \in \mathcal{P}_0$. 

Proof of 1st main theorem

Proof.

- If $\mathcal{P}_0 = \emptyset$ then, for all $n \geq 1$, we have

$$d(P_n, Q_n) = d(\mathcal{P}_0, Q) = \inf_{P \in \mathcal{P}_0, Q \in Q} d(P, Q) = \inf \emptyset = \infty$$  \hspace{1cm} (21)

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 (18) holds for some $P \in \mathcal{P}_0$.

- Then, lemma 2.3 ($c + b_{n-1} \geq b_n + a_n$, $n = 1, 2, \ldots$) with

$$c = d(P, Q), \quad b_n = d(P, Q_n), \quad a_n = d(P_n, Q_n)$$  \hspace{1cm} (22)
Proof of 1st main theorem

Proof.

- If $P_0 = \emptyset$ then, for all $n \geq 1$, we have

  \[ d(P_n, Q_n) = d(P_0, Q) = \inf_{P \in P_0, Q \in Q} d(P, Q) = \inf \emptyset = \infty \quad (21) \]

  so theorem is true in this case (l.h.s. holds by definition of $P_0$).

- Suppose $P_0 \neq \emptyset$ and that Eq (18) holds for some $P \in P_0$.

- Then, lemma 2.3 ($c + b_{n-1} \geq b_n + a_n$, $n = 1, 2, \ldots$) with

  \[ c = d(P, Q), \quad b_n = d(P, Q_n), \quad a_n = d(P_n, Q_n) \quad (22) \]

  Why? Since $P \in P_0$, we have both $\exists n_0$ s.t., $b_{n_0} < \infty$ and also $c < \infty$ (all by the def of $P_0$) implying $a_n < \infty$ for $n \geq n_0$. 

Prof. Jeff Bilmes
Proof of 1st main theorem

Proof.

- If $P_0 = \emptyset$ then, for all $n \geq 1$, we have

$$d(P_n, Q_n) = d(P_0, Q) = \inf_{P \in P_0, Q \in Q} d(P, Q) = \inf \emptyset = \infty \quad (21)$$

so theorem is true in this case (l.h.s. holds by definition of $P_0$).

- Suppose $P_0 \neq \emptyset$ and that Eq (18) holds for some $P \in P_0$.

- Then, lemma 2.3 ($c + b_{n-1} \geq b_n + a_n$, $n = 1, 2, \ldots$) with

$$c = d(P, Q), \quad b_n = d(P, Q_n), \quad a_n = d(P_n, Q_n) \quad (22)$$

Why? Since $P \in P_0$, we have both $\exists n_0$ s.t., $b_{n_0} < \infty$ and also $c < \infty$ (all by the def of $P_0$) implying $a_n < \infty$ for $n \geq n_0$.

- Also, $\limsup_{n \to \infty} b_n > -\infty$ since Eq (18) 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 2.3 holds here and from it we get:
Proof of 1st main theorem

Proof.

- So, lemma 2.3 holds here and from it we get:
- Under A: \( \forall P \in P_0, \) we have

\[
\liminf_{n \to \infty} a_n = \liminf_{n \to \infty} d(P_n, Q_n) \leq c = d(P, Q) < \infty \quad (23)
\]
Proof. 

So, lemma 2.3 holds here and from it we get:

Under A: \( \forall P \in \mathcal{P}_0 \), we have

\[
\liminf_{n \to \infty} a_n = \liminf_{n \to \infty} d(P_n, Q_n) \leq c = d(P, Q) < \infty \tag{23}
\]

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 \( \liminf_{n \to \infty} d(P_n, Q_n) \leq d(\mathcal{P}_0, Q) < \infty \)).
Proof of 1st main theorem

Proof.

- So, lemma 2.3 holds here and from it we get:

- Under A: \( \forall P \in \mathcal{P}_0 \), we have

\[
\liminf_{n \to \infty} a_n = \liminf_{n \to \infty} d(P_n, Q_n) \leq c = d(P, Q) < \infty \tag{23}
\]

- 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 \( \{P_n, Q_n\} \in \mathcal{P}_0 \).
Proof.

- So, lemma 2.3 holds here and from it we get:

- Under A: \( \forall P \in \mathcal{P}_0 \), we have

\[
\liminf_{n \to \infty} a_n = \liminf_{n \to \infty} d(P_n, Q_n) \leq c = d(P, Q) < \infty \quad (23)
\]

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

\((\{P_n, Q_n\})_{n=0}^{\infty}\)

Also, let \( P = P_{n-1} \) in Eq (18), so we get:

\[
d(P_{n-1}, Q) + d(P_{n-1}, Q_{n-1}) \geq d(P_{n-1}, Q_n) + d(P_n, Q_n) \quad (24)
\]

\[
\ldots
\]
Proof of 1st main theorem

Proof.

This implies that

\[
d(P_{n-1}, Q_{n-1}) \geq d(P_{n-1}, Q_{n}) - d(P_{n-1}, Q) + d(P_n, Q_n) \geq d(P_n, Q_n)
\]
Proof.

This implies that

\[ d(P_{n-1}, Q_{n-1}) \geq d(P_{n-1}, Q_n) - d(P_{n-1}, Q) + d(P_n, Q_n) \geq 0 \]

\[ \geq 0 \]

(25)

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.
Proof.

This implies that

\[
d(P_{n-1}, Q_{n-1}) \geq d(P_{n-1}, Q_n) - d(P_{n-1}, Q) + d(P_n, Q_n) \geq d(P_n, Q_n) \geq 0
\]

(25)

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

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

(26)

\ldots
Proof.

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

$$d(P_n, Q_n) \geq d(\mathcal{P}_0, Q) = d(P, Q) \quad (27)$$

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

Proof.

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

$$d(P_n, Q_n) \geq d(\mathcal{P}_0, Q) = d(P, Q)$$  \hspace{1cm} (27)

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

This means,

$$a_n = d(P_n, Q_n) \geq c = d(P, Q)$$  \hspace{1cm} (28)

or that $c - a_n \leq 0$, implying that $\sum_{n=0}^{\infty} (c - a_n)^+ < \infty$
Proof of 1st main theorem

Proof.

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

$$d(P_n, Q_n) \geq d(\mathcal{P}_0, Q) = d(P, Q) \tag{27}$$

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

This means,

$$a_n = d(P_n, Q_n) \geq c = d(P, Q) \tag{28}$$

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

From lemma 2.3, this gives $\lim_{n \to \infty} a_n = c$ or

$$\lim_{n \to \infty} d(P_n, Q_n) = d(P, Q) = d(\mathcal{P}_0, Q) \tag{29}$$
Proof of 1st main theorem

Proof.

- And since (as shown earlier)

\[ \sum_{n=n_1}^{\infty} (a_n - c) < \infty \]  \hfill (30)
Proof of 1st main theorem

Proof.

- And since (as shown earlier)

\[ \sum_{n=n_1}^{\infty} (a_n - c) < \infty \tag{30} \]

- we have

\[ \sum_{n=n_1}^{\infty} (d(P_n, Q_n) - d(P_0, Q)) < \infty \tag{31} \]

\[ \tag{32} \]

\[ \tag{33} \]
Sequences

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

\[
P_0 \xrightarrow{1} Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \xrightarrow{2} P_2 \xrightarrow{1} Q_2 \xrightarrow{2} P_3 \xrightarrow{1} Q_3 \xrightarrow{2} \cdots \quad (34)
\]
Consider next sequences \( \{(P_n, Q_n)\}_{n=0}^{\infty} \) constructed by alternating minimization with arbitrary starting point \( P_0 \in \mathcal{P} \)

\[
P_0 \xrightarrow{1} Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \xrightarrow{2} P_2 \xrightarrow{1} Q_2 \xrightarrow{2} P_3 \xrightarrow{1} Q_3 \xrightarrow{2} \cdots \tag{34}
\]

Then we have that:

\[
d(P_n, Q_n) \geq d(P_{n+1}, Q_n) \geq d(P_{n+1}, Q_{n+1}) \quad \text{for } n = 0, 1, \ldots \tag{35}
\]
Sequences

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

\[
P_0 \xrightarrow{1} Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \xrightarrow{2} P_2 \xrightarrow{1} Q_2 \xrightarrow{2} P_3 \xrightarrow{1} Q_3 \xrightarrow{2} \cdots \quad (34)
\]

- Then we have that:

\[
d(P_n, Q_n) \geq d(P_{n+1}, Q_n) \geq d(P_{n+1}, Q_{n+1}) \quad \text{for } n = 0, 1, \ldots \quad (35)
\]

- 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 $P_0 \in \mathcal{P}$

\[
P_0 \xrightarrow{1} Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \xrightarrow{2} P_2 \xrightarrow{1} Q_2 \xrightarrow{2} P_3 \xrightarrow{1} Q_3 \xrightarrow{2} \cdots	ag{34}
\]

- Then we have that:

\[
d(P_n, Q_n) \geq d(P_{n+1}, Q_n) \geq d(P_{n+1}, Q_{n+1}) \quad \text{for } n = 0, 1, \ldots \tag{35}
\]

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

- If 5PP holds for some $P \in \mathcal{P}$ and if we construct an alternating minimization sequence starting at some $P_0 \in \mathcal{P}$, we have conditions of Theorem 2.4 met at $P \in \mathcal{P}_0$
Sequences

So Theorem 2.4 holds in this case (i.e.,
\[ \lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \].
Sequences

- So Theorem 2.4 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.
So Theorem 2.4 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 2.4 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.
So Theorem 2.4 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 2.5 (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 \( \forall Q_0 \)

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

(36)
Definition 2.5 (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 \( \forall Q_0 \)

\[
\delta(P, P_1) + d(P_1, Q_0) \leq d(P, Q_0) \quad \text{whenever} \quad Q_0 \xrightarrow{2} P_1 \tag{36}
\]

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

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

\[ P \in \mathcal{P} \]

\[ P_1 \in \arg\min_{P \in \mathcal{P}} d(P, Q_0) \]
Definition 2.6 (Four Points Property (4PP))

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

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

(37)
Four Points Property (4PP)

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

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

Theorem 3.1

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

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(\mathcal{P}_0, Q) \tag{38}
\]

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

**Theorem 3.1**

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

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \tag{38}
\]

if \( P \) (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. \( 3PP + 4PP \Rightarrow 5PP \)
Theorem 3.1

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

\[
\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q) \tag{38}
\]

if \( P \) (defined by either: A) all \( P \in \mathcal{P}_0 \); or B) some \( P \in \mathcal{P}_0 \) with \( d(P, Q) = d(\mathcal{P}_0, Q) \) 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 a \( P \) for which A holds.
Second Main Theorem

Proof.

- We saw already: 5PP + alternating minimization ⇒ Theorem 2.4.
Second Main Theorem

Proof.

- We saw already: 5PP + alternating minimization $\Rightarrow$ Theorem 2.4.
- Combining 3PP and 4PP we have that when: $Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1$

\[
d(P, Q_0) - \delta(P, P_1) \geq d(P_1, Q_0) \quad 3PP \tag{39}
\]
\[
\delta(P, P_1) + d(P, Q) \geq d(P, Q_1) \quad 4PP \tag{40}
\]

...
Second Main Theorem

Proof.

- We saw already: 5PP + alternating minimization \( \Rightarrow \) Theorem 2.4.
- Combining 3PP and 4PP we have that when: \( Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \)

\[
d(P, Q_0) - \delta(P, P_1) \geq d(P_1, Q_0) \quad \text{3PP} \tag{39}
\]
\[
\delta(P, P_1) + d(P, Q) \geq d(P, Q_1) \quad \text{4PP} \tag{40}
\]

- 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 already: 5PP + alternating minimization $\Rightarrow$ Theorem 2.4.
- Combining 3PP and 4PP we have that when: $Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1$

\[
\begin{align*}
    d(P, Q_0) - \delta(P, P_1) &\geq d(P_1, Q_0) \quad \text{3PP} \\
    \delta(P, P_1) + d(P, Q) &\geq d(P, Q_1) \quad \text{4PP}
\end{align*}
\]

- 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:

\[
\begin{align*}
    d(P, Q_0) + d(P, Q) &\geq d(P, Q_1) + d(P_1, Q_0) \quad (41) \\
    &\geq d(P, Q_1) + d(P_1, Q_1) \quad (42)
\end{align*}
\]

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

Proof.

Further, if both of 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, with \( P_n \xrightarrow{1} Q_n \xrightarrow{2} P_{n+1} \) (via 4PP and 3PP resp.)

\[
\delta(P, P_{n+1}) + d(P_{n+1}, Q_n) \leq d(P, Q_n) \leq \delta(P, P_n) + d(P, Q) \quad (43)
\]
Second Main Theorem

Proof.

Further, if both of 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, with \( P_n \xrightarrow{1} Q_n \xrightarrow{2} P_{n+1} \) (via 4PP and 3PP resp.)

\[ \delta(P, P_{n+1}) + d(P_{n+1}, Q_n) \leq d(P, Q_n) \leq \delta(P, P_n) + d(P, Q) \tag{43} \]

This implies

\[ \delta(P, P_{n+1}) \leq \delta(P, P_n) + [d(P, Q) - d(P_{n+1}, Q_n)] \forall Q \tag{44} \]

So

\[ \delta(P, P_{n+1}) \leq \delta(P, P_n) + \underbrace{[d(P, Q) - d(P_{n+1}, Q_n)]}_{\leq 0} \tag{45} \]
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:

$$\lim_{n \to \infty} d(P_n, Q_n) = d(P_0, Q)$$ (46)
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} (46)

- 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:

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

(46)

- 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} (46)

- 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 soon 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

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

(47)

is such 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., $\mathbb{R}^n$
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 $\delta(P, P') = \|P - P'\|^2$. 
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 $\geq \pi/2$. 

\[
\begin{align*}
\phi & \geq \pi/2 \\
\text{if } & \text{ and } \mathcal{P} \text{ are both convex.}
\end{align*}
\]
Example: Squared Euclidean-Distance

- Let $\mathcal{P}$ and $\mathcal{Q}$ be two non-empty closed convex subsets of $\mathbb{R}^N$. 
Example: Squared Euclidean-Distance

- Let $\mathcal{P}$ and $\mathcal{Q}$ be two non-empty closed convex subsets of $\mathbb{R}^N$.
- Let $d(P, Q) = \| P - Q \|_2^2$ and $\delta(P, P') = \| P - P' \|_2^2$
Example: Squared Euclidean-Distance

- Let $P$ and $Q$ be two non-empty closed convex subsets of $\mathcal{R}^N$.
- Let $d(P, Q) = \|P - Q\|_2^2$ and $\delta(P, P') = \|P - P'\|_2^2$
- Goal is to find $\min_{P \in P, P \in Q} d(P, Q)$
Example: Squared Euclidean-Distance

- Let $\mathcal{P}$ and $\mathcal{Q}$ be two non-empty closed convex subsets of $\mathbb{R}^N$.
- Let $d(P, Q) = \|P - Q\|_2^2$ and $\delta(P, P') = \|P - P'\|_2^2$.
- 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:

$$
\|P - Q_0\|_2^2 = \|P - P_1 + P_1 - Q_0\|_2^2 \tag{48}
$$

$$= \|P - P_1\|_2^2 + \|P_1 - Q_0\|_2^2 + 2 \langle P - P_1, P_1 - Q_0 \rangle \tag{49}
$$

This gives the result since $\langle P - P_1, P_1 - Q_0 \rangle \geq 0$. 

Prof. Jeff Bilmes
page 34
**Example: Squared Euclidean-Distance**

- Let $\mathcal{P}$ and $\mathcal{Q}$ be two non-empty closed convex subsets of $\mathbb{R}^N$.
- Let $d(P, Q) = \|P - Q\|_2^2$ and $\delta(P, P') = \|P - P'\|_2^2$
- 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:

\[
\|P - Q_0\|_2^2 = \|P - P_1 + P_1 - Q_0\|_2^2
\]
\[
= \|P - P_1\|_2^2 + \|P_1 - Q_0\|_2^2 + 2 \langle P - P_1, P_1 - Q_0 \rangle
\]
Example: Squared Euclidean-Distance

- Let $\mathcal{P}$ and $\mathcal{Q}$ be two non-empty closed convex subsets of $\mathbb{R}^N$.
- Let $d(P, Q) = \|P - Q\|_2^2$ and $\delta(P, P') = \|P - P'\|_2^2$
- 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:

$$
\|P - Q_0\|_2^2 = \|P - P_1 + P_1 - Q_0\|_2^2
$$

$$
= \|P - P_1\|_2^2 + \|P_1 - Q_0\|_2^2 + 2 \langle P - P_1, P_1 - Q_0 \rangle
$$

(48)

(49)

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

- Geometry of why $\langle P - P_1, P_1 - Q_0 \rangle \geq 0$
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) $\mathcal{P}$
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) $\mathcal{P}$

Exercise: Prove 4PP for squared Euclidean-Distance.
Example

- This also satisfies 4PP since angle at $Q_1$ are $\geq \pi/2$ (exercise: prove this).
Example

- This also satisfies 4PP since angle at $Q_1$ are $\geq \pi/2$ (exercise: prove this).

- 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).
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 \(P \in \mathcal{P}\).
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 \(\mathcal{P}, Q\) be convex (e.g., simplex or convex subsets thereof)
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 measureable 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

\[
d(P, Q) = D(P \| Q) = \begin{cases} \int \log p dP & \text{if } P \ll Q \\ \infty & \text{if } P \not\ll Q \end{cases} \tag{50}
\]

where \(p = \frac{dP}{dQ}\) is the Radon-Nikodym derivative (more later)
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

$$d(P, Q) = D(P \| Q) = \begin{cases} \int \log p \, dP & \text{if } P \ll Q \\ \infty & \text{if } P \not\ll Q \end{cases} \quad (50)$$

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).
• \( p = \frac{dP}{dQ} \) is the Radon-Nikodym derivative, meaning that \( \forall E \in \mathcal{X} \), we have that \( p \) is defined so that

\[
p(E) = \int_E \left[ \frac{dP}{dQ} \right] dQ \tag{51}
\]

I.e., we see \( \frac{dP}{dQ} \) as a function that maps one measure to the other.
Convex Sets

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

$$p(E) = \int_E \left[ \frac{dP}{dQ} \right] dQ$$  \hspace{1cm} (51)

i.e., we see $\frac{dP}{dQ}$ as a function that maps one measure to the other.

- Thus, if $P \ll Q$, then

$$D(P||Q) = \int \log \left[ \frac{dP}{dQ} \right] dP = \int p(x) \log \frac{p(x)}{q(x)} dx$$  \hspace{1cm} (52)
p = \frac{dP}{dQ} is the Radon-Nikodym derivative, meaning that \forall E \in \mathcal{X}, we have that p is defined so that

\[ p(E) = \int_E \left[ \frac{dP}{dQ} \right] dQ \]  \hspace{1cm} (51)

I.e., we see \frac{dP}{dQ} as a function that maps one measure to the other.

Thus, if \( P \ll Q \), then

\[ D(P||Q) = \int \log \left[ \frac{dP}{dQ} \right] dP = \int p(x) \log \frac{p(x)}{q(x)} dx \] \hspace{1cm} (52)

And if \( P \not\ll Q \), then, what happens?
Convex Sets

For $P, P' \in \mathcal{P}$, we have the generalized KL as follows:

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

(53)
Convex Sets

- For $P, P' \in \mathcal{P}$, we have the generalized KL as follows:

  $$\delta(P, P') \triangleq D(P\|P') + P'(X) - P(X) \geq 0$$  \hspace{1cm} (53)

- So if the measures are probability measures then $P'(\mathcal{X}) = P(\mathcal{X}) = 1$, and we get back standard KL-divergence.
Convex Sets

For $P, P' \in \mathcal{P}$, we have the generalized KL as follows:

$$\delta(P, P') \triangleq D(P||P') + P'(X) - P(X) \geq 0 \quad (53)$$

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 4.1

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 $Q_0 \xrightarrow{1} P_1$, then we have

$$D(P||P_1) + P_1(X) - P(X) + D(P_1||Q_0) \leq D(P||Q_0) \tag{54}$$

Proof.

- By definition, $D(P_1||Q_0) = D(\mathcal{P}||Q_0) < \infty$. 

...
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 $Q_0 \xrightarrow{1} P_1$, then we have

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

(54)

Proof.

- By definition, $D(P_1||Q_0) = D(\mathcal{P}||Q_0) < \infty$.
- Assume $D(P||Q_0) < \infty$ since otherwise we get the inequality immediately.
Theorem 4.1

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 \( Q_0 \overset{1}{\rightarrow} P_1 \), then we have

\[
D(P || P_1) + P_1(X) - P(X) + D(P_1 || Q_0) \leq D(P || Q_0)
\]  

(54)

Proof.

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

\[
p_1 = \frac{dP_1}{dQ_0}, \quad \text{and} \quad p = \frac{dP}{dQ_0}
\]

(55)

so that \( P_1 = \int p_1 dQ_0 \) and \( P = \int p dQ_0 \)
3PP KL

Proof.

Form $P_\alpha = (1 - \alpha)P + \alpha P_1$ as a convex combination of $P$ and $P_1$.
3PP KL

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

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

(56)

since \( Q_0 \xrightarrow{2} P_1 \).
3PP KL

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

$$f(1) = D(P_1 \| Q_0) < D(P_\alpha \| Q_0) = f(\alpha) \quad (56)$$

since $Q_0 \xrightarrow{2} P_1$.

- Thus,

$$0 \geq \frac{f(1) - f(\alpha)}{1 - \alpha} = \frac{1}{1 - \alpha} \left[ \int dP_1 \log \frac{dP_1}{dQ_0} - \int dP_\alpha \log \frac{dP_\alpha}{dQ_0} \right] \quad (57)$$

$$= \frac{1}{1 - \alpha} \left[ \int p_1 dQ_0 \log \frac{dP_1}{dQ_0} - \int p_\alpha dQ_0 \log \frac{dP_\alpha}{dQ_0} \right] \quad (58)$$

$$= \frac{1}{1 - \alpha} \left[ \int (p_1 \log p_1 - p_\alpha \log p_\alpha) dQ_0 \right] \quad (59)$$
Proof.

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

- 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 $\alpha \uparrow 1$. 

$$\lim_{\alpha \to 1} \left( \text{quantity} \right) = \frac{d}{d\alpha} \left( \text{quantity} \right) \bigg|_{\alpha=1} \quad (60)$$

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

$$0 \geq \int \frac{d}{d\alpha} \left( p_\alpha \log p_\alpha \right) \bigg|_{\alpha=1} dQ_0 = \int \left( 1 + \log p_1 \right) \left( p_1 - p \right) dQ_0 \quad (61)$$

$$= \int p_1 dQ_0 - p dQ_0 + p_1 \log p_1 dQ_0 - p \log p_1 dQ_0 \quad (62)$$

From this, the results follow using definitions of $p_1$ and $p$. . .
Proof.

- 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,

\[
\lim_{\alpha \uparrow 1} (\text{quantity}) = \left. \frac{d}{d\alpha} (\text{quantity}) \right|_{\alpha=1} \tag{60}
\]

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

\[
0 \geq \int \left. \frac{d}{d\alpha} (p_\alpha \log p_\alpha) \right|_{\alpha=1} dQ_0 = \int (1 + \log p_1) (p_1 - p_0) dQ_0 \tag{61}
\]

\[
= \int p_1 dQ_0 - p_0 dQ_0 + p_1 \log p_1 dQ_0 - p \log p_1 dQ_0 \tag{62}
\]

From this, the results follow using definitions of \( p_1 \) and \( p_0 \) ...
Proof.

- 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,

$$\lim_{\alpha \uparrow 1} (\text{quantity}) = \frac{d}{d\alpha} (\text{quantity}) \Bigg|_{\alpha=1}$$

(60)

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

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

(61)

$$= \int p_1 dQ_0 - p dQ_0 + p_1 \log p_1 dQ_0 - p \log p_1 dQ_0$$

(62)
Proof.

- 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,

$$\lim_{\alpha \uparrow 1} (\text{quantity}) = \left. \frac{d}{d\alpha} (\text{quantity}) \right|_{\alpha=1}$$

(60)

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

$$0 \geq \int \left. \frac{d}{d\alpha} (p_\alpha \log p_\alpha) \right|_{\alpha=1} dQ_0 = \int (1 + \log p_1)(p_1 - p) dQ_0$$

(61)

$$= \int p_1 dQ_0 - p dQ_0 + p_1 \log p_1 dQ_0 - p \log p_1 dQ_0$$

(62)

- From this, the results follow using definitions of $p_1$ and $p$.
Theorem 4.2

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

$$D(P \| Q_1) \leq D(P \| P_1) + P_1(X) - P(X) + D(P \| Q)$$  \hspace{1cm} (63)

$$d(P, Q_1) \leq \delta(P, P_1) + d(P, Q)$$  \hspace{1cm} (64)

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 4.3**

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

$$\lim_{n \to \infty} D(P_n || Q_n) = D(P_0 || Q) \quad (65)$$

where

$$\mathcal{P}_0 = \{ P : D(P || Q_n) < \infty \text{ for some } n \} \quad (66)$$

Also, if $X$ is a finite set (+ a few other minor technical conditions) then $P_n \to P^*$ where $D(P^* || Q) = D(P_0 || Q)$. 

Prof. Jeff Bilmes
Moreover, if $X$ is finite, and $P_0$ is positive for $x \in X$ such that
\[ \exists P \in \mathcal{P}, Q \in \mathcal{Q} \text{ with } P(x)Q(x) > 0 \text{ (simultaneously positive on } X), \]
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).
A main theorem (continued)

Moreover, if $X$ is finite, and $P_0$ is positive for $x \in X$ such that
\[ \exists P \in \mathcal{P}, Q \in \mathcal{Q} \text{ with } P(x)Q(x) > 0 \] (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.
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?
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 $D(P_n||Q_n) - D(\mathcal{P}_0||Q)$?
A main theorem (continued)

Moreover, if $X$ is finite, and $P_0$ is positive for $x \in X$ such that
\[ \exists P \in \mathcal{P}, Q \in \mathcal{Q} \text{ with } P(x)Q(x) > 0 \] (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 $D(P \| Q) = D(\mathcal{P}_0 \| Q)$ then from 5PP we get

\[
D(\mathcal{P}_0 \| Q) + D(P \| Q_{n-1}) \geq D(P \| Q_n) + D(P_n \| Q_n)
\] (67)
This implies that

\[ D(P_n || Q_n) - D(P_0 || Q) \leq D(P || Q_{n-1}) - D(P || Q_n) \]  
\[ = \int \log \frac{dQ_n}{dQ_{n-1}} dP \]  
\[ \leq \log \sup_x \frac{dQ_n(x)}{dQ_{n-1}(x)} \to 0 \]

if the sequence has a limit (i.e., is convergent) so that it does not change from iteration to iteration.
This implies that

\[ D(P_n \| Q_n) - D(P_0 \| Q) \leq D(P \| Q_{n-1}) - D(P \| Q_n) \]  \hspace{1cm} (68)

\[ = \int \log \frac{dQ_n}{dQ_{n-1}} dP \]  \hspace{1cm} (69)

\[ \leq \log \sup_x \frac{dQ_n(x)}{dQ_{n-1}(x)} \rightarrow 0 \]  \hspace{1cm} (70)

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 at ratios of successive measures.
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).
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.
Scratch Paper