1. **Review:** [Ch. 4 of "Markov Chains and Mixing Times" by Peres et al.]

- **Total Variation Distance:** \(\Omega - \) finite alphabet, \(P - \) simplex of pmfs on \(\Omega\)
  
  \[\forall \mu, \nu \in P, \quad \|\mu - \nu\|_{TV} = \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|\]

  \[= \frac{1}{2} \|\mu - \nu\|_1\]

  \[= \sum_{x \in \Omega} |\mu(x) - \nu(x)|\]

  \[= \frac{1}{2} \max_{f: \Omega \to \mathbb{R}} \text{subject to } \max_{x \in \Omega} |f(x)| \leq 1\]

  \[\left\{\begin{array}{l}
  \text{variational/functional characterization} \\
  \text{analogue to dual representation of Wasserstein distance, where} \\
  \text{the constraint is } \text{Lip}(f) \leq 1.
\end{array}\right\}\]

  \[\text{optimal coupling representation}\]

  \[\text{pictorial interpretation}\]

- **Convergence Thm:** Given irreducible and aperiodic MC \(P\) with stationary pmf \(\pi\),
  
  \[\exists \alpha(0, 1)\text{ and } C > 0 \text{ s.t. } \max_{x \in \Omega} \|P_t(x) - \pi\|_{TV} \leq C \alpha^t.\]

  **Remark:** Two important proofs due to Doeblin \(\to\) minorization [Ch. 4],

  coupling [Ch. 5].

- **Ergodic Thm:** For any \(f: \Omega \to \mathbb{R}\) and an irreducible MC \(\{X_t\}\), we have:
  
  \[\forall \mu \in P, \quad \mathbb{P}\left(\lim_{t \to \infty} \frac{1}{t} \sum_{s=0}^{t-1} f(X_s) = \mathbb{E}_\pi[f]\right) = 1\]

  where \(\pi\) is the stationary pmf.

  **Remark:** Proof uses SLLN after partitioning MC into stopping time intervals.

- **Contraction Properties:** MC \(P\) with stationary dist. \(\pi\)
  
  \[d(t) \leq \max_{x \in \Omega} \|P_t(x) - \pi\|_{TV}\]

  \[\bar{d}(t) \leq \max_{x, y \in \Omega} \|P_t(x, \cdot) - P_t(y, \cdot)\|_{TV} \leq \bar{d}(t)\]

  \[\text{is the contraction coefficient for TV distance}\]

  - \(d(t) \leq \bar{d}(t) \leq 2d(t)\)
  - \(d(s+t) \leq d(s)d(t)\) [submultiplicative property]
  - \(d(t), \bar{d}(t)\) non-increasing in \(t\) [trivial from DPI]
Exercises: [4.2 is real analysis, 4.4 is obvious from 4.3, 4.3 will be done more generally.]

4.1 Prop: 0 \pi \in \Omega \Rightarrow \max_{x \in \Omega} \|x \pi - \pi\|_\pi = \sup_{\mu \in P} \| \mu \pi - \pi\|_\pi
\Rightarrow \max_{x,y \in \Omega} \|x \pi - y \pi\|_\pi = \sup_{\mu \nu \in \Omega} \| \mu \nu - \pi \nu\|_\pi
\Rightarrow \max_{\mu \nu \in \Omega} \| \mu \nu - \pi \nu\|_\pi = \sup_{\mu \nu \in \Omega} \| \mu \nu - \pi \nu\|_\pi

Pf: [\pi] follows because \pi is compact and \mu \mapsto \| \mu \pi - \pi\|_\pi, (\mu, \nu) \mapsto \| \mu \nu - \pi \nu\|_\pi
are continuous functions \Rightarrow use Extreme Value Theorem.

1. (\leq) Obvious.
2. (\geq) Obvious.

4.5 Prop: Let \mu_i and \nu_i be measures on \Omega_i (finite set) for i = 1, ..., n.

Define \mu = \prod_{i=1}^n \mu_i, \nu = \prod_{i=1}^n \nu_i on \prod_{i=1}^n \Omega_i.

Then, \| \mu - \nu \|_\pi \leq \sum_{i=1}^n \| \mu_i - \nu_i \|_\pi.

Pf: Let (x_i, y_i) be the optimal coupling of \mu_i and \nu_i st. P_{x_i} = \mu_i and P_{x_i} = \nu_i.

Then, \| \mu_i - \nu_i \|_\pi = \mathbb{P}(x_i \neq y_i) for i = 1, ..., n.

Let (x, y) be independent for i = 1, ..., n, and let x = \prod_{i=1}^n x_i, y = \prod_{i=1}^n y_i.

(x, y) is a coupling of \mu and \nu because P_x = \mu and P_y = \nu.

Then, \| \mu - \nu \|_\pi \leq \mathbb{P}(x \neq y) = \mathbb{P}(\exists i st. x_i \neq y_i) \leq \sum_{i=1}^n \mathbb{P}(x_i \neq y_i) = \sum_{i=1}^n \| \mu_i - \nu_i \|_\pi.
2. Coefficients of Ergodicity:

- introduced in the context of convergence rates of finite inhomogeneous MCs
- ergodicity: long-term behaviour of products of stochastic matrices

**Weak Ergodicity:** inhomogeneous MC

Let \( S_{k} \) be a sequence of \( n \times n \) row stochastic matrices, and \( T^{(p, r)} = \prod_{i=1}^{r} S_{p+i} \).

**Def:** (Kolmogorov) \( S_{k} \) is weakly ergodic if \( \forall i, j, s \in \{1, \ldots, n\} \) and \( p > 0 \),

\[
\lim_{r \to \infty} T^{(p, r)}_{i j} - T^{(p, r)}_{i s} = 0.
\]

**Remark:** As \( r \to \infty \), rows of product converge and become independent of initial pmf. Note that \( T^{(p, r)}_{i j} \) does not necessarily tend to a limit; \( T^{(p, r)} \) is rank 1 for larger \( p \) depends on \( r \).

**Remark:** If in addition, \( \forall i, j, s \in \{1, \ldots, n\}, p > 0 \), \( \lim_{r \to \infty} T^{(p, r)}_{i j} \) exists, then \( S_{k} \) is strongly ergodic. (Also, all rows tend to some \( \pi \), and \( \exists \pi \text{ s.t. } T^{(p, r)}_{i j} \to \pi \forall i, j, r \to \infty \).

**Contraction Coefficient:**

**Def:** A coefficient of ergodicity \( \eta() \) is a continuous function from stochastic matrices to \([0, 1]\). Such a coefficient is proper if \( \eta(S) = 0 \iff S = 1_p \)

for some pmf \( p \) (or equivalently, \( \text{rank}(S) = 1 \)).

**Thm:** \( S_{k} \) is weakly ergodic if and only if \( \forall p > 0, \lim_{r \to \infty} \eta(T^{(p, r)}) = 0 \),

where \( \eta() \) is a proper coefficient of ergodicity.

**Pf:** \( \Rightarrow \) \( S_{k} \) weakly ergodic \( \iff \) \( T^{(p, r)} \) becomes rank 1 as \( r \to \infty \) (but may not be fixed), \( \forall p > 0 \)

\( \Rightarrow \) \( \eta(T^{(p, r)}) \to 0 \) as \( r \to \infty \), by continuity of \( \eta() \).

\( \Leftarrow \)

Suppose \( \forall p > 0, \lim_{r \to \infty} \eta(T^{(p, r)}) = 0 \) and \( S_{k} \) is not weakly ergodic.

**Observe:** Let \( C = \{ M \in \text{stochastic} | M = 1_p \text{ for some pmf } p \} \).

Then,

\[
\{ S_{k} \} \text{ weakly ergodic } \iff \forall p > 0, \lim_{r \to \infty} \inf_{M \in C} \| M - T^{(p, r)} \|_{F_0} = 0.
\]

Hence, \( \exists \{ D_{m} \} \text{ subseq. of } S_{k}, D_{m} > 0 \) s.t.

\[
\inf_{M \in C} \| M - T^{(p, r)} \|_{F_0} > \epsilon \text{ for some } p > 0.
\]

Since stochastic matrices are compact, \( T^{(p, r)} \to p^{*} \text{-stochastic} \) [where we may use a subsequence of \( \{ D_{m} \} \) if necessary by Bolzano-Weierstrass Thm].

So, \( \eta(T^{(p, r)}) \to \eta(p^{*}) \) as \( r \to \infty \) [continuity of \( \eta() \)], and \( \eta(T^{(p, r)}) \to 0 \) as \( r \to \infty \)

by assumption. Hence, \( \eta(p^{*}) = 0 \) and \( p^{*} \in C \), which is a contradiction.

**Remark:** If \( S_{k} \) is homogeneous with \( S_{k} = S \), then it is weakly ergodic

if and only if \( \lim_{r \to \infty} \eta(S^{r}) = 0 \). Note: If \( \eta(S) < \eta(S)^{n} \text{ (submultiplicative)} \),

then such convergence is easy to prove.
**Information Theoretic Examples:**

Def: (Csiszar, Morimoto, Ali-Silvey) Given a convex function \( f: \mathbb{R}^+ \to \mathbb{R} \) s.t. \( f(1) = 0 \) and \( f(\cdot) \) is strictly convex at \( t = 1 \) (i.e. \( \forall x, y \) s.t. \( x + y = 1 \) for any \( x \leq y ) \), \( f(x) < f(x(1 + x)) \),

\[ D_f(\mu \| \nu) = \sum_{x \in \Omega} \nu(x) \frac{f(\frac{\nu(x)}{\mu(x)})}{f'(\frac{1}{\mu(x)})} \]

is the \( f \)-divergence between \( \mu \) and \( \nu \).

**Remark:** \( f(0) = \lim_{t \to 0^+} f(t) = 0 \), \( of(t) = 0 \), \( of'(t) = r \lim_{s \to 0^+} sf'(t) \), \( \forall r > 0 \).

**Examples:**
1. \( f(t) = t \log(t) \) \( \to \) KL divergence
2. \( f(t) = t^2 - 1 \) \( \to \) \( \chi^2 \)-divergence
3. \( f(t) = \frac{1}{2} |t - 1| \) \( \to \) total variation distance

**Properties:**
1. **[Non-negativity]** \( \forall \mu, \nu \in \mathcal{P}, \ D_f(\mu \| \nu) \geq 0 \) with equality iff \( \mu = \nu \).
2. **[Joint Convexity]** \( (\mu, \nu) \mapsto D_f(\mu \| \nu) \) is jointly convex.
3. **[Data Processing Inequality]** \( \forall \mu, \nu \in \mathcal{P}, \ D_f(\mu \| \nu) \leq D_f(\mu \| \nu') \) for stochastic matrix \( P \).

**Exercise 43**

\( f(t) = \frac{1}{2} |t - 1| \)

**Lemma:** (Perspective Function) \( f: \mathbb{R}^+ \to \mathbb{R} \), \( f(x) \) convex \( \iff (p, q) \mapsto qf(\frac{p}{q}) \) convex.

**Proof:**

\( \iff (\leftarrow) \) set \( q = 1 \).

\( \leftarrow \)Fix \( \lambda \in [0, 1] \), \( \lambda \mu = \lambda x - \lambda y \). Observe that:

\[ (\lambda q_1 + \lambda q_2, f\left(\frac{\lambda q_1 + \lambda q_2}{\lambda q_1 + \lambda q_2}\right) = (\lambda q_1 + \lambda q_2) f\left(\frac{\lambda q_1 + \lambda q_2}{\lambda q_1 + \lambda q_2}\right) \leq \lambda q_1 f\left(\frac{p_1}{q_1}\right) + \lambda q_2 f\left(\frac{p_2}{q_2}\right) \]

[\( f \) convex].

1. \( \sum_{x \in \Omega} \nu(x) f\left(\frac{\nu(x)}{\mu(x)}\right) \geq f\left(\sum_{x \in \Omega} \nu(x)\right) = 0 \) with equality iff \( \mu = \nu \).

2. Obvious from Lemma.

3. Fix \( y \in \Omega \) and let \( \sum_{x \in \Omega} P(x, y) \). Observe that:

\[ \sum_{x \in \Omega} \frac{\nu(x) P(x, y)}{Z(y)} f\left(\frac{\sum_{x \in \Omega} \nu(x) P(x, y)}{Z(y)}\right) \leq \sum_{x \in \Omega} \frac{P(x, y)}{Z(y)} \nu(x) f\left(\frac{\nu(x)}{\mu(x)}\right) \]

[from Lemma]

\[ \Rightarrow \sum_{y \in \Omega} \frac{(\mu P)(y)}{Z(y)} f\left(\frac{(\mu P)(y)}{(\mu P)(y)}\right) \leq \sum_{x \in \Omega} \nu(x) f\left(\frac{\mu(x)}{\mu(x)}\right) \]

\[ D_f(\mu \| \nu) \]

\[ D_f(\mu \| \nu') \]

\[ \sum_{x \in \Omega} \nu(x) f\left(\frac{\nu(x)}{\mu(x)}\right) \leq \sum_{x \in \Omega} \nu(x) f\left(\frac{\nu(x)}{\mu(x)}\right) \]
Def: (Contraction Coefficient) For a stochastic matrix $P$, we define:

$$
\eta_f(P) \equiv \sup_{\mu, \nu \in P} \frac{D_f(\mu P \| \nu P)}{D_f(\mu \| \nu)}
$$

This gives strong data processing inequalities: $\forall \mu, \nu \in P$, $D_f(\mu P \| \nu P) \leq \eta_f(P) D_f(\mu \| \nu)$. 

Thm: $\eta_f(\cdot)$ satisfies the following:

1. $0 \leq \eta_f(P) \leq 1$,

2. $P \mapsto \eta_f(P)$ is convex,

3. $P \mapsto \eta_f(P)$ is continuous on the interior of all stochastic matrices,

4. $\eta_f(P) = 0 \iff \text{rank}(P) = 1$. [

$\eta_f(\cdot)$ is a proper coefficient of ergodicity

Pf: 1. Obvious from DPI and non-negativity of $D_f(\| \cdot \| )$.

2. Fix $\mu, \nu \in P$ s.t. $0 < D_f(\mu \| \nu) < \infty$. Then, $P \mapsto \frac{D_f(\mu P \| \nu P)}{D_f(\mu \| \nu)}$ is jointly convex. Since $P \mapsto \eta_f(P)$ is a pointwise supremum of convex functionals on a convex compact set, it is convex.

3. Every convex function is continuous on the interior of its domain (use 2).

4. $(\Rightarrow) P = 1 \pi \Rightarrow \mu P = \nu P = \pi \Rightarrow D_f(\mu P \| \nu P) = 0, \forall \mu, \nu \in P \Rightarrow \eta_f(P) = 0$.

$(\Leftarrow) \eta_f(P) = 0 \Rightarrow \forall \mu, \nu \in P$ s.t. $0 < D_f(\mu \| \nu) < \infty$, $D_f(\mu P \| \nu P) = 0$

$\Rightarrow \forall \mu, \nu \in P$ s.t. $0 < D_f(\mu \| \nu) < \infty, \mu P = \nu P$

$\Rightarrow \forall \mu, \nu \in \text{relint}(P)$, $(\mu - \nu)P = 0$

(For any $v \perp 1$, $\exists c \neq 0$ s.t. $\mu + cv = \nu \in \text{relint}(P)$ where $\mu \in \text{relint}(P)$. So, $\forall v \perp 1$, $c \neq 0$. Subtract $1$ row vector $v=c(\mu - \nu)$ for $\mu, \nu \in \text{relint}(P)$.)

$\Rightarrow \forall v \perp 1, \nu P = 0$, i.e. leftnull($P$) = $\{v \in \mathbb{R}^n : v^T = 0\}$ and nullity($P$) = $n - 1$.

3. Doeblin Minorization: [See proof of Convergence Thm in Ch. 4.]

- Doeblin minorization condition: A Markov matrix $P$ satisfies the minorization condition if $\exists \theta \in (0, 1), \exists \pi \in P$ s.t. $P \geq \theta \pi \pi^T$ entrywise. $\theta$ a minorization constant.

Thm: If $P$ satisfies Doeblin minorization, then $\eta_f(P) \leq \theta$.

Pf: $P$ satisfies minorization $\Rightarrow \hat{P} \equiv P - \theta \pi \pi^T$ is a valid stochastic matrix. Let $E_\theta$ denote the stochastic matrix of an erasure channel with prob. $\theta$ of erasure.

Then, $P = \theta E_\theta + (1 - \theta) \pi \pi^T = E_\theta T$.

Observe that $\forall \mu, \nu \in P$, $D_f(\mu P \| \nu P) = D_f(\mu E_\theta P \| \nu E_\theta P) \leq D_f(\mu E_\theta \| \nu E_\theta)$.
Dobrushin Contraction Coefficient:

- $\eta_f(P)$ for $f(t) = \frac{1}{2} |t-1|$ is the contraction coefficient for total variation distance.

**Def. (Dobrushin Coefficient)** For a MC $P$, $\eta_{TV}(P) \leq \sup_{\mu \neq \nu} \frac{||\mu P - \nu P||_{TV}}{||\mu - \nu||_{TV}}$. Can replace with $l_1$-norm.

**Prop:** (Various Representations) $\eta_{TV}(P) = \max_{v : ||v||_1 = 1} ||vP||_1 = \max_{x, y : \mu x = \nu y} ||P(x) - P(y)||_{TV}$ for any two rows $x, y$.

**Remark:** $\eta_{TV}(P) = I(t)$ [from Ch. 4].

**Thm. (Properties of $\eta_{TV}()$)**
- $\forall P$, $\eta_{TV}(P) \geq \eta_f(P)$ [Cohen, Kemperman, Zbaganu]
- (Lipschitz continuity) $||P_1 - P_2||_{TV} \leq ||P_1 - P_2||_{l_1}$
- (Subdominant Eigenvalue Bound) $|\lambda| \leq \eta_{TV}(P)$ for all eigenvalues $\lambda \neq 1$ of $P$ [Bauer, Deutsch, Stoer]
- (Submultiplicative Property) $\eta_{TV}(P_1P_2) \leq \eta_{TV}(P_1)\eta_{TV}(P_2)$ [Dobrushin] generalizes sub-mult prop of $I(t)$ in Ch. 4.

**Pf:** 2 WLOG, let $\eta_{TV}(P) > \eta_{TV}(P_2)$. Also, let $\eta_{TV}(P) = ||vP||_1$ for some $v \perp 1$, $||v||_1 = 1$.

$0 \leq ||vP||_1 - \max_{x \perp 1} ||xP||_1 \leq ||vP||_1 - ||vP_2||_1 \leq ||P(x) - P_2(x)||_1 = ||(P - P_2)v||_1 \leq ||(P - P_2)||_{TV}$ row sum.

3 (Real subdominant eigenvalue case) If $\lambda \neq 1$ is an eigenvalue of $P$, then $xP = \lambda x$ for some row vector $x$. Since $P1 = 11$, $x \perp 1$ (left and right-eigs resp. to distinct vals are 1).

Let $||x||_1 = 1$. Then, we have:

$|\lambda| = |\lambda||x||_1 = ||xP||_1 \leq \max_{v : ||v||_1 = 1} ||vP||_1 = \eta_{TV}(P)$.

4 Let $\eta_{TV}(P_1P_2) = ||xP_1P_2||_1$ for some row vector $x$ s.t. $||x||_1$ and $x \perp 1$.

Let $y = \frac{xP_1}{||xP_1||}$, $||y||_1 = 1$ and $y1 = \frac{1}{||xP_1||} xP_1 |xP_1| = 0$, i.e., $y \perp 1$.

$\Rightarrow \eta_{TV}(P_1P_2) = ||xP_1P_2||_1 = ||xP_1||_1 ||yP_2||_1 = ||xP_1||_1 ||yP_2||_1 \leq \eta_{TV}(P_1)\eta_{TV}(P_2)$.

**Remark:** 2 and 3 show why $\eta_{TV}(\cdot)$ is useful. The second largest eigenvalue modulus (SLEM) controls the rate of convergence to stationarity, but it is not sub-multiplicative. $\eta_{TV}(\cdot)$ bounds SLEM and allows convergence analysis as it is sub-multiplicative.

References:

1. "Markov Chains and Mixing Times" by Levin, Peres, and Wilmer [Ch. 4].
2. "Stochastic Matrices: Ergodicity Coefficients, and Applications to Ranking" by S.T. Margaret [Ch. 3].
3. "Non-negative Matrices and Markov Chains" by Seneta [Ch. 3 & 4].
5. "Strong Beta Processing Inequalities and $\ell_1$-Sobolev Inequalities for Discrete Channels" by Raginsky.