Anuran Makur
8 June 2016

Introduction to Markov Mixing:

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

- Total Variation Distance: Ω - finite alphabet, $\mathcal{P}$ - simplex of pmfs on Ω

$$\forall \mu, \nu \in \mathcal{P}, \quad \|\mu - \nu\|_{TV} \triangleq \max_{A \subseteq \Omega} |\mu(A) - \nu(A)|$$

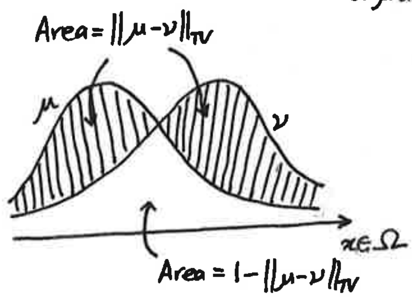
$$= \tfrac{1}{2}\|\mu - \nu\|_1 \quad \Big\} \ \ell_1\text{-norm characterization}$$

$$= \sum_{\substack{x \in \Omega: \\ \mu(x) \geq \nu(x)}} \mu(x) - \nu(x)$$

$$= \frac{1}{2} \max_{\substack{f: \Omega \to \mathbb{R} \\ \|f\|_\infty \triangleq \max_{x\in\Omega} |f(x)| \leq 1}} \mathbb{E}_\mu[f] - \mathbb{E}_\nu[f] \quad \Big\} \text{ variational/functional characterization}$$

- analogous to dual representation of 1-Wasserstein distance, where the constraint is $Lip(f) \leq 1$.

$$= \min_{(X,Y) \text{ coupling of } \mu \text{ and } \nu} \mathbb{P}(X \neq Y) \quad \Big\} \text{ optimal coupling representation}$$



} pictorial interpretation

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

  Remark: Two important proofs due to Doeblin → minorization [Ch. 4], → coupling [Ch. 5].

- Ergodic Thm: For any $f: \Omega \to \mathbb{R}$ and an irreducible MC $\{X_t\}$, we have:

$$\forall \mu \in \mathcal{P}, \quad \mathbb{P}_\mu\left(\lim_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} f(X_s) = \mathbb{E}_\pi[f]\right) = 1$$

  where π is the stationary pmf.

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

- Contraction Properties: MC P with stationary dist. π

$$d(t) \triangleq \max_{x\in\Omega} \|P^t(x,\cdot) - \pi\|_{TV}$$

$$\bar{d}(t) \triangleq \max_{x,y\in\Omega} \|P^t(x,\cdot) - P^t(y,\cdot)\|_{TV} \quad \leftarrow \bar{d}(1) \text{ is the \underline{contraction coefficient} for TV distance}$$

  - $d(t) \leq \bar{d}(t) \leq 2d(t)$
  - $\bar{d}(s+t) \leq \bar{d}(s)\bar{d}(t)$ [submultiplicative property]
  - $d(t), \bar{d}(t)$ non-increasing in t [trivial from DPI]

Anuran Makur

- Exercises: [4.2 is real analysis, 4.4 is obvious from 4.3, 4.3 will be done more generally]

4.1 Prop: ① $d(t) \triangleq \max_{x\in\Omega} \|\delta_x P^t - \pi\|_{TV} = \sup_{\mu\in\mathcal{P}} \|\mu P^t - \pi\|_{TV} \overset{[\star]}{=} \max_{\mu\in\mathcal{P}} \|\mu P^t - \pi\|_{TV}$

② $\bar{d}(t) \triangleq \max_{x,y\in\Omega} \|\delta_x P^t - \delta_y P^t\|_{TV} = \sup_{\mu,\nu\in\mathcal{P}} \|\mu P^t - \nu P^t\|_{TV} \overset{[\star]}{=} \max_{\mu,\nu\in\mathcal{P}} \|\mu P^t - \nu P^t\|_{TV}$

Pf: $[\star]$ follows because $\mathcal{P}$ is compact and $\mu \mapsto \|\mu P^t - \pi\|_{TV}$, $(\mu,\nu) \mapsto \|\mu P^t - \nu P^t\|_{TV}$ are continuous functions $\Rightarrow$ use Extreme Value Theorem.

① $(\leq)$ Obvious.

$(\geq)$ $\max_{\mu\in\mathcal{P}} \|\mu P^t - \pi\|_{TV} = \max_{\mu\in\mathcal{P}} \left\|\sum_{x\in\Omega} \mu(x)\delta_x P^t - \mu(x)\pi\right\|_{TV} \underset{\text{ineq.}}{\overset{\text{Triangle}}{\leq}} \max_{\mu\in\mathcal{P}} \sum_{x\in\Omega} \mu(x) \|\delta_x P^t - \pi\|_{TV}$

$\leq \max_{x\in\Omega} \|\delta_x P^t - \pi\|_{TV}$.

② $(\leq)$ Obvious.

$(\geq)$ $\forall \nu\in\mathcal{P}$, $\max_{\mu\in\mathcal{P}} \|\mu P^t - \nu P^t\|_{TV} \leq \max_{x\in\Omega} \|\delta_x P^t - \nu P^t\|_{TV}$ [previous proof]

$\forall x\in\Omega$, $\max_{\nu\in\mathcal{P}} \|\delta_x P^t - \nu P^t\|_{TV} \leq \max_{y\in\Omega} \|\delta_x P^t - \delta_y P^t\|_{TV}$ [previous proof]

$\Rightarrow \max_{\mu,\nu\in\mathcal{P}} \|\mu P^t - \nu P^t\|_{TV} \leq \max_{x\in\Omega,\, \nu\in\mathcal{P}} \|\delta_x P^t - \nu P^t\|_{TV} \leq \max_{x,y\in\Omega} \|\delta_x P^t - \delta_y P^t\|_{TV}$. ∎

4.5 Prop: Let μ_i and ν_i be measures on Ω_i (finite set) for $i=1,\dots,n$.

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

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

Pf: Let (X_i, Y_i) be the optimal coupling of μ_i and ν_i s.t. $P_{X_i} = \mu_i$ and $P_{Y_i} = \nu_i$.

Then, $\|\mu_i - \nu_i\|_{TV} = \mathbb{P}(X_i \neq Y_i)$ for $i=1,\dots,n$.

Let (X_i, Y_i) be independent for $i=1,\dots,n$, and let $X = X_1^n$, $Y = Y_1^n$.

(X,Y) is a coupling of μ and ν because $P_X = \mu$ and $P_Y = \nu$.

Then, $\|\mu - \nu\|_{TV} \underset{\substack{\text{coupling} \\ \text{characterization}}}{\overset{TV}{\leq}} \mathbb{P}(X \neq Y) = \mathbb{P}(\exists i \text{ s.t. } X_i \neq Y_i) \underset{\text{bound}}{\overset{\text{union}}{\leq}} \sum_{i=1}^{n} \mathbb{P}(X_i \neq Y_i) = \sum_{i=1}^{n} \|\mu_i - \nu_i\|_{TV}$.

∎

Anuran Makur

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

Def: (Kolmogorov) $\{S_k\}$ is weakly ergodic if $\forall\, i,j,s \in \{1,\dots,n\}$ and $p \geq 0$,

$$\lim_{r\to\infty} T_{is}^{(p,r)} - T_{js}^{(p,r)} = 0.$$

Remark: As no. of factors $\to \infty$, rows of product equalize and become indep. of initial pmf. Note that $T_{is}^{(p,r)}$ does not necessarily tend to a limit; $T^{(p,r)}$ is $\approx$ rank 1 for large r, but $T^{(p,r)}$ deps on r.

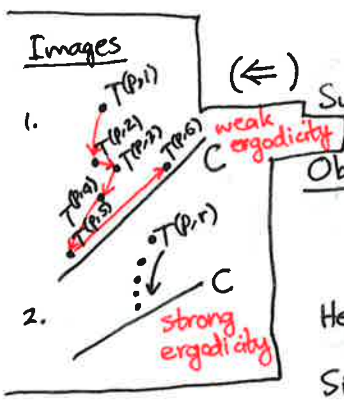
Remark: If in addition, $\forall i,s \in \{1,\dots,n\}$, $p \geq 0$, $\lim_{r\to\infty} T_{is}^{(p,r)}$ exists, then $\{S_k\}_{k=1}^{\infty}$ is strongly ergodic. (Also, all rows tend to some π, and $\exists p$ s.t. $T^{(p,r)} \to \underset{\uparrow\,\text{all ones col. vector}}{1}\pi \Leftrightarrow \forall p\geq 0, T^{(p,r)} \to 1\pi$

- **Contraction Coefficient:**

Def: A coefficient of ergodicity $\eta(\cdot)$ is a continuous function from stochastic matrices to $[0,1]$. Such a coefficient is proper if $\eta(S) = 0 \Leftrightarrow S = 1p$ (S: $n\times n$ stoch. matrix) for some pmf p (or equivalently, $\text{rank}(S) = 1$).

Thm: $\{S_k\}_{k=1}^{\infty}$ is weakly ergodic if and only if $\forall p \geq 0$, $\lim_{r\to\infty} \eta(T^{(p,r)}) = 0$, where $\eta(\cdot)$ is a proper coefficient of ergodicity.

Pf: ($\Rightarrow$) $\{S_k\}$ weakly ergodic $\Leftrightarrow$ $T^{(p,r)}$ becomes rank 1 as $r\to\infty$ (but may not be fixed), $\forall p \geq 0$
$\Rightarrow \eta(T^{(p,r)}) \to 0$ as $r\to\infty$, by continuity of $\eta(\cdot)$.



($\Leftarrow$) Suppose $\forall p \geq 0$, $\lim_{r\to\infty} \eta(T^{(p,r)}) = 0$ and $\{S_k\}_{k=1}^{\infty}$ is not weakly ergodic.

Observe: Let $C = \{M\ n\times n \text{ stochastic} \mid M = 1p \text{ for some pmf } p\}$. Then,
$\{S_k\}$ weakly ergodic $\Leftrightarrow$ $\forall p \geq 0$, $\lim_{r\to\infty} \underbrace{\inf_{M\in C} \|M - T^{(p,r)}\|_{Fro}}_{=\text{distance}(T^{(p,r)},\,C)} = 0$.

Hence, $\exists \{r_m\}$ subseq. of $\{r\}$, $\exists \varepsilon > 0$ s.t. $\inf_{M\in C} \|M - T^{(p,r_m)}\|_{Fro} > \varepsilon$, $\forall m$ [for some $p \geq 0$].

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

So, $\eta(T^{(p,r_m)}) \to \eta(P^*)$ as $r_m\to\infty$ [continuity of η], and $\eta(T^{(p,r_m)}) \to 0$ as $r_m \to \infty$ by assumptiom. 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 $\lim_{r\to\infty} \eta(\underbrace{S^r}_{T^{(0,r)}}) = 0$. Note: If $\eta(S^r) \leq \eta(S)^r$ [submultiplicative], then such convergence is easy to prove.

Anuran Makur

- Information Theoretic Examples:

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

$\forall \mu, \nu \in \mathcal{P}$, $D_f(\mu \| \nu) \triangleq \sum_{x \in \Omega} \nu(x) f\left(\frac{\mu(x)}{\nu(x)}\right)$ is the f-divergence between μ and ν.

Remark: $f(0) = \lim_{t \to 0^+} f(t)$, $0 f\left(\frac{0}{0}\right) = 0$, $0 f\left(\frac{r}{0}\right) = \lim_{s \to 0^+} s f\left(\frac{r}{s}\right) = r \lim_{s \to 0^+} s f\left(\frac{1}{s}\right)$, $\forall r > 0$.

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.

Exercise 43 → $f(t) = \frac{1}{2}|t-1|$

③ [Data Processing Inequality] $\forall \mu, \nu \in \mathcal{P}$, $D_f(\mu P \| \nu P) \leq D_f(\mu \| \nu)$ for stochastic matrix P.

Proof: Lemma: (Perspective Function) $f: \mathbb{R} \to \mathbb{R}$, $f(p)$ convex $\Leftrightarrow$ $(p,q) \mapsto q f\left(\frac{p}{q}\right)$ convex. $p \in \mathbb{R}, q \in \mathbb{R}^+$

Pf: ($\Leftarrow$) Set $q = 1$.

($\Rightarrow$) Fix $\lambda \in [0,1]$, $\bar{\lambda} \triangleq 1 - \lambda$. Observe that:

$$(\lambda q_1 + \bar{\lambda} q_2) f\left(\frac{\lambda p_1 + \bar{\lambda} p_2}{\lambda q_1 + \bar{\lambda} q_2}\right) = (\lambda q_1 + \bar{\lambda} q_2) f\left(\frac{\lambda q_1}{\lambda q_1 + \bar{\lambda} q_2} \cdot \frac{p_1}{q_1} + \frac{\bar{\lambda} q_2}{\lambda q_1 + \bar{\lambda} q_2} \cdot \frac{p_2}{q_2}\right)$$

$$\leq \lambda q_1 f\left(\frac{p_1}{q_1}\right) + \bar{\lambda} q_2 f\left(\frac{p_2}{q_2}\right) \quad [f \text{ convex}]. \quad ▨$$

① $$\sum_{x \in \Omega} \nu(x) f\left(\frac{\mu(x)}{\nu(x)}\right) \overset{\text{Jensen's}}{\underset{\text{inequality}}{\geq}} f\left(\sum_{x \in \Omega} \mu(x)\right) = 0$$ with equality iff $\mu = \nu$.

② Obvious from Lemma.

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

$$\underbrace{\sum_{x \in \Omega} \nu(x) \frac{P(x,y)}{Z(y)}}_{\frac{(\nu P)(y)}{Z(y)}} f\left(\frac{\sum_{x \in \Omega} \mu(x) \frac{P(x,y)}{Z(y)}}{\sum_{x \in \Omega} \nu(x) \frac{P(x,y)}{Z(y)}}\right) \leq \sum_{x \in \Omega} \frac{P(x,y)}{Z(y)} \nu(x) f\left(\frac{\mu(x)}{\nu(x)}\right) \quad [\text{from Lemma}]$$

$\sum_{y \in \Omega}$ on both sides

$$\Rightarrow \underbrace{\sum_{y \in \Omega} (\nu P)(y) f\left(\frac{(\mu P)(y)}{(\nu P)(y)}\right)}_{D_f(\mu P \| \nu P)} \leq \underbrace{\sum_{x \in \Omega} \nu(x) f\left(\frac{\mu(x)}{\nu(x)}\right)}_{D_f(\mu \| \nu)} \quad ▨$$

Anuran Makur

Def: (Contraction Coefficient) For a stochastic matrix P, we define:

$$\eta_f(P) \triangleq \sup_{\substack{\mu,\nu \in \mathcal{P} \\ 0 < D_f(\mu\|\nu) < \infty}} \frac{D_f(\mu P \| \nu P)}{D_f(\mu\|\nu)} .$$

This gives strong data processing inequalities: $\forall \mu,\nu \in \mathcal{P}$, $\overbrace{D_f(\mu P\|\nu P) \leq \eta_f(P)\, D_f(\mu\|\nu)}^{\text{measure ergodicity wrt } D_f(\cdot\|\cdot)}$.

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

① $0 \leq \eta_f(P) \leq 1$, $\star$

② $P \mapsto \eta_f(P)$ is convex,

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

④ $\eta_f(P) = 0 \Leftrightarrow \text{rank}(P) = 1$. $\star$ $[\Leftrightarrow P = 1\pi$ for some $\pi \in \mathcal{P}]$ corresponds to independence of X and Y if $P_{Y|X}=P$

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

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

② Fix $\mu,\nu \in \mathcal{P}$ s.t. $0 < D_f(\mu\|\nu) < \infty$. Then, $P \mapsto \dfrac{D_f(\mu P\|\nu P)}{D_f(\mu\|\nu)}$ is convex in P as $D_f(\cdot\|\cdot)$ 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 $(\Leftarrow)$ $P = 1\pi \Rightarrow \mu P = \nu P = \pi \Rightarrow D_f(\mu P\|\nu P) = 0$, $\forall \mu,\nu \in \mathcal{P} \Rightarrow \eta_f(P) = 0$.

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

$\Leftrightarrow \forall \mu,\nu \in \mathcal{P}$ s.t. $0 < D_f(\mu\|\nu) < \infty$, $\mu P = \nu P$

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

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

$\Rightarrow \forall v \perp 1$, $vP = 0$, i.e. $\text{leftnull}(P) = \{v \in \mathbb{R}^n : \underbrace{v}_{\text{row vector}} 1 = 0\}$ and $\text{nullity}(P) = n-1$ ∎

$\Rightarrow \text{rank}(P) = 1$.

---

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

• Doeblin minorization condition: A Markov matrix P satisfies the minorization condition if $\exists \theta \in (0,1)$, $\exists \pi \in \mathcal{P}$ s.t. $P \geq \bar{\theta}\, \underset{\text{col. vec}}{1}\, \underset{\text{row vec.}}{\pi}$ entrywise. $[\bar{\theta} \triangleq 1 - \theta]$

Thm: If P satisfies Doeblin minorization, then $\eta_f(P) \leq \underset{\text{minorization constant}}{\theta}$.

Pf: P satisfies minorization $\Rightarrow \hat{P} \triangleq \dfrac{P - \bar{\theta} 1 \pi}{\theta}$ is a valid stochastic matrix.

Let E_θ denote the stochastic matrix of an erasure channel with prob. $\bar{\theta}$ of erasure.

Then, $$P = \underbrace{\begin{array}{c|c} & \overbrace{\quad\Omega\quad}^{} \;\; \overbrace{e}^{\text{erasure}} \\ \Omega & \begin{bmatrix} \theta & & 0 & \bar{\theta} \\ & \ddots & & \vdots \\ 0 & & \theta & \bar{\theta} \end{bmatrix} \end{array}}_{E_\theta} \cdot \underbrace{\begin{array}{c|c} & \Omega \\ \Omega & \begin{bmatrix} \hat{P} \\ \hline \pi \end{bmatrix} \\ e & \end{array}}_{T} = E_\theta T.$$

Observe that: $\forall \mu,\nu \in \mathcal{P}$, $D_f(\mu P\|\nu P) = D_f(\mu E_\theta T \| \nu E_\theta T) \overset{\text{DPI}}{\leq} D_f(\mu E_\theta \| \nu E_\theta) = D_f(\underbrace{\theta\mu + \bar{\theta}\delta_e}_{\text{pmfs on } \Omega \cup \{e\}} \| \underbrace{\theta\nu + \bar{\theta}\delta_e}) \overset{\text{convexity}}{\leq} \theta D_f(\mu\|\nu)$ ∎

Anuran Makur

④ 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, $\eta_{TV}(P) \triangleq \sup_{\substack{\mu,\nu \in \mathcal{P} \\ \mu \neq \nu}} \frac{\|\mu P - \nu P\|_{TV}}{\|\mu - \nu\|_{TV}}$. ← can replace with ℓ_1-norm

Prop: (Various Representations)
$$\eta_{TV}(P) = \underbrace{\max_{\substack{v: \|v\|_1 = 1 \\ v^T \perp \mathbf{1}}} \|vP\|_1}_{\text{Dobrushin}} = \underbrace{\max_{x,y \in \Omega} \|P(x,\cdot) - P(y,\cdot)\|_{TV}}_{\text{Markov-Doeblin}} = \underbrace{1 - \min_{x,y\in\Omega} \sum_{z\in\Omega} \min\{P_{xz}, P_{yz}\}}_{\text{Dobrushin}}.$$

↑ row vector

$\overset{[\text{Ex.4.1}]}{=} \max_{\mu,\nu \in \mathcal{P}} \|\mu P - \nu P\|_{TV}$

>0 for scrambling matrices ($\eta_{TV} < 1$)
* for any two rows ∃ col. s.t. the probs are >0

Remark: $\eta_{TV}(P^t) = \bar{d}(t)$ [from Ch. 4].

Thm: (Properties of $\eta_{TV}(\cdot)$)

① $\forall P$, $\eta_{TV}(P) \geq \eta_f(P)$ [Cohen, Kempermann, Zbăganu]

② (Lipschitz continuity) $\forall P_1, P_2$, $|\eta_{TV}(P_1) - \eta_{TV}(P_2)| \leq \|P_1 - P_2\|_\infty$

③ (Subdominant Eigenvalue Bound) $|\lambda| \leq \eta_{TV}(P)$ for all eigenvalues $\lambda \neq 1$ of P [Bauer, Deutsch, Stoer]

④ (Submultiplicative Property) $\eta_{TV}(P_1P_2) \leq \eta_{TV}(P_1)\eta_{TV}(P_2)$ [Dobrushin] ← generalizes sub-mult prop of $\bar{d}(\cdot)$ in Ch. 4.

Pf: ② WLOG let $\eta_{TV}(P_1) \geq \eta_{TV}(P_2)$. Also, let $\eta_{TV}(P_1) = \|vP_1\|_1$ for some $v^T \perp \mathbf{1}$, $\|v\|_1 = 1$.
$$\Rightarrow 0 \leq \underbrace{\|vP_1\|_1}_{\eta_{TV}(P_1)} - \underbrace{\max_{\substack{z: \|z\|_1 = 1 \\ z^T\perp \mathbf{1}}} \|zP_2\|_1}_{\eta_{TV}(P_2)} \leq \|vP_1\|_1 - \underbrace{\|vP_2\|_1}_{\|vP_2\|_1 \leq \eta_{TV}(P_2)} \leq \|v(P_1 - P_2)\|_1 = \|(P_1^T - P_2^T)v^T\|_1 \leq \underset{\substack{\uparrow \\ \text{max abs.} \\ \text{col sum}}}{\|P_1^T - P_2^T\|_1} = \underset{\substack{\uparrow \\ \text{max abs.} \\ \text{row sum}}}{\|P_1 - P_2\|_\infty}.$$

③ (Real subdominant eigenvalue case) If $\lambda \neq 1$ is an eigenvalue of P, then $xP = \lambda x$ for some row vector x. Since $P\mathbf{1} = 1.\mathbf{1}$, $x^T \perp \mathbf{1}$ (left and right e-vecs corresp. to distinct e-vals are $\perp$). Let $\|x\|_1 = 1$. Then, we have:
$$|\lambda| = |\lambda|\|x\|_1 = \|xP\|_1 \leq \max_{\substack{v: \|v\|_1 = 1 \\ v^T\perp\mathbf{1}}} \|vP\|_1 = \eta_{TV}(P).$$

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

Let $y = \frac{xP_1}{\|xP_1\|_1} \Rightarrow \|y\|_1 = 1$ and $y\mathbf{1} = \frac{1}{\|xP_1\|_1} xP_1\mathbf{1} = \frac{x\mathbf{1}}{\|xP_1\|_1} = 0$, i.e. $y^T \perp \mathbf{1}$.
$$\Rightarrow \eta_{TV}(P_1P_2) = \|xP_1P_2\|_1 = \big\| \|xP_1\|_1 \, yP_2 \big\|_1 = \|xP_1\|_1 \|yP_2\|_1 \leq \eta_{TV}(P_1)\eta_{TV}(P_2).$$ ▨

Remark: ① and ② 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.

⑤ 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]
4. "Coefficients of Ergodicity: Structure and Applications" by Seneta.
5. "Strong Data Processing Inequalities and Φ-Sobolev Inequalities for Discrete Channels" by Raginsky.