# Geometrical Aspects of Data-Processing of Markov Chains

Beyond IID in Information Theory 10

arXiv:2203.04575 *Geoffrey Wolfer*†, *Shun Watanabe*‡

September 26, 2022

† RIKEN AIP – Special Postdoctoral Researcher Program
‡ Tokyo University of Agriculture and Technology

# Introduction & Preliminaries

# Information geometry

Finite space $\mathcal{Y} \cong [m]$. Simplex over $\mathcal{Y}$: $\mathcal{P}(\mathcal{Y})$.

# Information geometry

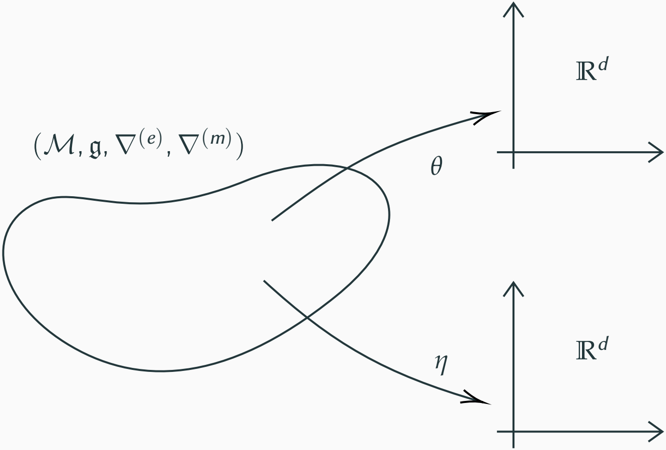
Finite space $\mathcal{Y} \cong [m]$. Simplex over $\mathcal{Y}$: $\mathcal{P}(\mathcal{Y})$.

View family $\mathcal{M} = \{\mu_\theta : \theta \in \Theta\} \subset \mathcal{P}(\mathcal{Y})$ as a smooth manifold.

# Information geometry

Finite space $\mathcal{Y} \cong [m]$. Simplex over $\mathcal{Y}$: $\mathcal{P}(\mathcal{Y})$.

View family $\mathcal{M} = \{\mu_\theta \colon \theta \in \Theta\} \subset \mathcal{P}(\mathcal{Y})$ as a smooth manifold.



# Information geometry

**Fisher information metric $\mathfrak{g}$ and dual affine connections $\nabla^{(e)}, \nabla^{(m)}$**

$$\mathfrak{g}_{ij}(\mu_\theta) \triangleq \sum_{y\in\mathcal{Y}} \mu_\theta(y)\partial_i \log \mu_\theta(y)\partial_j \log \mu_\theta(y),$$

$$\Gamma^{(e)}_{ij,k}(\mu_\theta) \triangleq \sum_{y\in\mathcal{Y}} \partial_i\partial_j \log \mu_\theta(y)\partial_k \mu_\theta(y),$$

$$\Gamma^{(m)}_{ij,k}(\mu_\theta) \triangleq \sum_{y\in\mathcal{Y}} \partial_i\partial_j \mu_\theta(y)\partial_k \log \mu_\theta(y).$$

# Information geometry

**Fisher information metric $\mathfrak{g}$ and dual affine connections $\nabla^{(e)}, \nabla^{(m)}$**

$$\mathfrak{g}_{ij}(\mu_\theta) \triangleq \sum_{y\in\mathcal{Y}} \mu_\theta(y)\partial_i \log \mu_\theta(y)\partial_j \log \mu_\theta(y),$$
$$\Gamma^{(e)}_{ij,k}(\mu_\theta) \triangleq \sum_{y\in\mathcal{Y}} \partial_i\partial_j \log \mu_\theta(y)\partial_k\mu_\theta(y),$$
$$\Gamma^{(m)}_{ij,k}(\mu_\theta) \triangleq \sum_{y\in\mathcal{Y}} \partial_i\partial_j\mu_\theta(y)\partial_k \log \mu_\theta(y).$$

**Wide range of applications**

1. Higher-order efficiency analysis of estimator.
2. Information decomposition / projection.
3. Natural gradient algorithms.

# Information geometry

**Fisher information metric $\mathfrak{g}$ and dual affine connections $\nabla^{(e)}, \nabla^{(m)}$**

$$\mathfrak{g}_{ij}(\mu_\theta) \triangleq \sum_{y \in \mathcal{Y}} \mu_\theta(y) \partial_i \log \mu_\theta(y) \partial_j \log \mu_\theta(y),$$

$$\Gamma^{(e)}_{ij,k}(\mu_\theta) \triangleq \sum_{y \in \mathcal{Y}} \partial_i \partial_j \log \mu_\theta(y) \partial_k \mu_\theta(y),$$

$$\Gamma^{(m)}_{ij,k}(\mu_\theta) \triangleq \sum_{y \in \mathcal{Y}} \partial_i \partial_j \mu_\theta(y) \partial_k \log \mu_\theta(y).$$

**Wide range of applications**

1. Higher-order efficiency analysis of estimator.
2. Information decomposition / projection.
3. Natural gradient algorithms.

**What about Markov models?**

# Irreducible Markov chains

**Notation**

$\mathcal{E} \subset \mathcal{Y}^2$ such that $(\mathcal{Y}, \mathcal{E})$ strongly connected.

Positive functions over $\mathcal{E}$: $\mathcal{F}_+(\mathcal{Y}, \mathcal{E})$.

Irreducible row-stochastic matrices over $(\mathcal{Y}, \mathcal{E})$: $\mathcal{W}(\mathcal{Y}, \mathcal{E})$.

# Irreducible Markov chains

**Notation**

$\mathcal{E} \subset \mathcal{Y}^2$ such that $(\mathcal{Y}, \mathcal{E})$ strongly connected.

Positive functions over $\mathcal{E}$: $\mathcal{F}_+(\mathcal{Y}, \mathcal{E})$.

Irreducible row-stochastic matrices over $(\mathcal{Y}, \mathcal{E})$: $\mathcal{W}(\mathcal{Y}, \mathcal{E})$.

Discrete-time, time-homogeneous Markov chain

$$\mathbb{P}\left(Y_1 = y_1, \dots, Y_k = y_k\right) = \mu(y_1) \prod_{t=1}^{k-1} P(y_t, y_{t+1}),$$

$$(\mu, P) \in (\mathcal{P}(\mathcal{Y}), \mathcal{W}(\mathcal{Y}, \mathcal{E})).$$

# Irreducible Markov chains

**Notation**

$\mathcal{E} \subset \mathcal{Y}^2$ such that $(\mathcal{Y}, \mathcal{E})$ strongly connected.

Positive functions over $\mathcal{E}$: $\mathcal{F}_+(\mathcal{Y}, \mathcal{E})$.

Irreducible row-stochastic matrices over $(\mathcal{Y}, \mathcal{E})$: $\mathcal{W}(\mathcal{Y}, \mathcal{E})$.

Discrete-time, time-homogeneous Markov chain

$$\mathbb{P}\left(Y_1 = y_1, \dots, Y_k = y_k\right) = \mu(y_1) \prod_{t=1}^{k-1} P(y_t, y_{t+1}),$$

$$(\mu, P) \in (\mathcal{P}(\mathcal{Y}), \mathcal{W}(\mathcal{Y}, \mathcal{E})).$$

Stationary distribution: $\pi P = \pi$.

Edge-measure: $Q(y, y') = \pi(y) P(y, y') = \mathbb{P}_\pi\left(Y_t = y, Y_{t+1} = y'\right)$.

# Exponential tilting (ET)

**ET of distribution**
$Y \sim \mu \in \mathcal{P}([m]), f \colon \mathcal{Y} \to \mathbb{R}$. Construct exponential family:

$$\mu_\theta(y) = \mu(y) e^{\theta f(y) - \kappa(\theta)}, \qquad \kappa(\theta) = \log \mathbb{E} e^{\theta f(Y)} \qquad \text{(CGF)}.$$

# Exponential tilting (ET)

**ET of distribution**
$Y \sim \mu \in \mathcal{P}([m]), f : \mathcal{Y} \to \mathbb{R}$. Construct exponential family:

$$\mu_\theta(y) = \mu(y) e^{\theta f(y) - \kappa(\theta)}, \qquad \kappa(\theta) = \log \mathbb{E} e^{\theta f(Y)} \qquad \text{(CGF)}.$$

$$\lambda > \mathbb{E}[f], \lim_{k \to \infty} -\frac{1}{k} \log \mathbb{P}\left( \frac{1}{k} \sum_{t=1}^{k} f(X_t) > \lambda \right) = \kappa^\star(\lambda) \triangleq \sup_{\theta \in \mathbb{R}} \{\theta\lambda - \kappa(\theta)\}.$$

# Exponential tilting (ET)

**ET of distribution**
$Y \sim \mu \in \mathcal{P}([m]), f\colon \mathcal{Y} \to \mathbb{R}$. Construct exponential family:

$$\mu_\theta(y) = \mu(y) e^{\theta f(y) - \kappa(\theta)}, \qquad \kappa(\theta) = \log \mathbb{E} e^{\theta f(Y)} \qquad \text{(CGF)}.$$

$$\lambda > \mathbb{E}[f], \lim_{k \to \infty} -\frac{1}{k} \log \mathbb{P}\left(\frac{1}{k}\sum_{t=1}^{k} f(X_t) > \lambda\right) = \kappa^\star(\lambda) \triangleq \sup_{\theta \in \mathbb{R}} \{\theta\lambda - \kappa(\theta)\}.$$

**ET of stochastic matrix**
$P \in \mathcal{W}, f\colon \mathcal{Y} \to \mathbb{R},\ \widetilde{P}_\theta(y, y') = P(y, y') \exp(\theta f(y')),$

# Exponential tilting (ET)

**ET of distribution**
$Y \sim \mu \in \mathcal{P}([m]), f\colon \mathcal{Y} \to \mathbb{R}$. Construct exponential family:

$$\mu_\theta(y) = \mu(y) e^{\theta f(y) - \kappa(\theta)}, \qquad \kappa(\theta) = \log \mathbb{E} e^{\theta f(Y)} \qquad \text{(CGF)}.$$

$$\lambda > \mathbb{E}\left[f\right], \lim_{k \to \infty} -\frac{1}{k} \log \mathbb{P}\left( \frac{1}{k} \sum_{t=1}^{k} f(X_t) > \lambda \right) = \kappa^\star(\lambda) \triangleq \sup_{\theta \in \mathbb{R}} \left\{ \theta \lambda - \kappa(\theta) \right\}.$$

**ET of stochastic matrix**
$P \in \mathcal{W}, f\colon \mathcal{Y} \to \mathbb{R},\ \widetilde{P}_\theta(y, y') = P(y, y') \exp(\theta f(y')),\ P_\theta = \mathfrak{s}(\widetilde{P}_\theta)$, with Perron-Frobenius (PF) rescaling (Miller, 1961),

$$\mathfrak{s}\colon \mathcal{F}_+(\mathcal{Y}, \mathcal{E}) \to \mathcal{W}(\mathcal{Y}, \mathcal{E}),\ \widetilde{P}(y, y') \mapsto P(y, y') = \frac{\widetilde{P}(y, y') v(y')}{\rho v(y)},$$

where $\rho$, $v$ are the PF root and right PF eigenvector of $\widetilde{P}$.

# Exponential tilting (ET)

**ET of distribution**
$Y \sim \mu \in \mathcal{P}([m]), f\colon \mathcal{Y} \to \mathbb{R}$. Construct exponential family:

$$\mu_\theta(y) = \mu(y) e^{\theta f(y) - \kappa(\theta)}, \qquad \kappa(\theta) = \log \mathbb{E} e^{\theta f(Y)} \qquad \text{(CGF)}.$$

$$\lambda > \mathbb{E}[f], \lim_{k\to\infty} -\frac{1}{k} \log \mathbb{P}\left(\frac{1}{k}\sum_{t=1}^{k} f(X_t) > \lambda\right) = \kappa^\star(\lambda) \triangleq \sup_{\theta\in\mathbb{R}} \{\theta\lambda - \kappa(\theta)\}.$$

**ET of stochastic matrix**
$P \in \mathcal{W}, f\colon \mathcal{Y} \to \mathbb{R}$, $\widetilde{P}_\theta(y, y') = P(y, y') \exp(\theta f(y'))$, $P_\theta = \mathfrak{s}(\widetilde{P}_\theta)$, with Perron-Frobenius (PF) rescaling (Miller, 1961),

$$\mathfrak{s}\colon \mathcal{F}_+(\mathcal{Y}, \mathcal{E}) \to \mathcal{W}(\mathcal{Y}, \mathcal{E}), \ \widetilde{P}(y, y') \mapsto P(y, y') = \frac{\widetilde{P}(y, y') v(y')}{\rho v(y)},$$

where $\rho$, $v$ are the PF root and right PF eigenvector of $\widetilde{P}$.

$$\lambda > \mathbb{E}[f], \lim_{k\to\infty} -\frac{1}{k} \log \mathbb{P}\left(\frac{1}{k}\sum_{t=1}^{k} f(X_t) > \lambda\right) = \sup_{\theta\in\mathbb{R}} \{\theta\lambda - \log\rho(\theta)\}.$$

# Information Geometry of Markov Chains

# Information Geometry of Markov Chains

1. Large deviations
   Miller (1961); Donsker and Varadhan (1975); Gärtner (1977).
2. Information projection
   Csiszár, Cover, and Choi (1987).
3. Asymptotic e-families
   Ito and Amari (1988); Takeuchi and Barron (1998); Takeuchi and Kawabata (2007); Takeuchi and Nagaoka (2017).
4. One-parameters exponential families
   Nakagawa and Kanaya (1993).
5. Dually flat structure
   Nagaoka (2005).

# Information Geometry of Markov Chains (Nagaoka, 2005)

Distributions

$\mathcal{P}(\mathcal{Y})$

$D$

KL divergence

$$D\left(\mu_\theta \| \mu_{\theta'}\right) = \mathbb{E}_{\mu_\theta} \log \frac{\mu_\theta(y)}{\mu_\theta(y')}$$

# Information Geometry of Markov Chains (Nagaoka, 2005)

Distributions

$\mathcal{P}(\mathcal{Y})$

Markov chains

$\mathcal{W}(\mathcal{Y}, \mathcal{E})$

$$Y_1, Y_2, \ldots, Y_t \sim P$$

$D \longrightarrow D$

KL divergence

KL divergence rate

$$\lim_{t\to\infty} \frac{1}{t} D\left( Y_1, \ldots, Y_t \sim P_\theta \middle\| Y'_1, \ldots, Y'_t \sim P_{\theta'} \right) = \mathbb{E}_{(Y,Y')\sim Q_\theta} \log \frac{P_\theta(Y,Y')}{P_{\theta'}(Y,Y')}$$

$$\triangleq D\left( P_\theta \middle\| P_{\theta'} \right)$$

# Information Geometry of Markov Chains (Nagaoka, 2005)

Distributions — $\mathcal{P}(\mathcal{Y})$

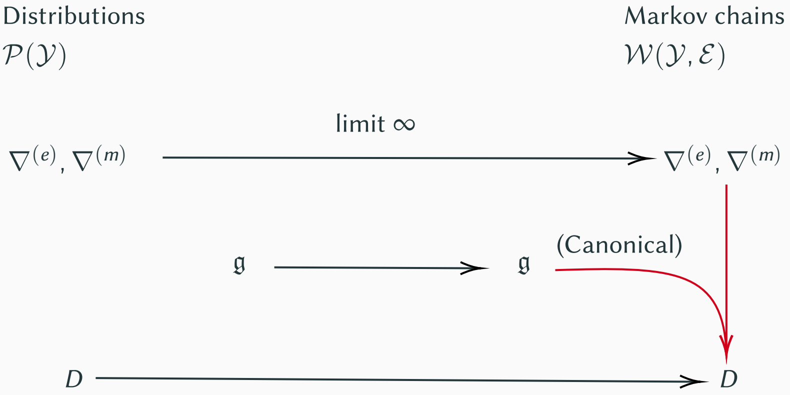
Markov chains — $\mathcal{W}(\mathcal{Y}, \mathcal{E})$

$$\nabla^{(e)}, \nabla^{(m)} \xrightarrow{\text{limit } \infty} \nabla^{(e)}, \nabla^{(m)}$$

$$\mathfrak{g} \longrightarrow \mathfrak{g}$$

$$D \longrightarrow D$$

# Information Geometry of Markov Chains (Nagaoka, 2005)



# Information geometry Markov chains (Nagaoka, 2005)

View $\mathcal{W}(\mathcal{Y}, \mathcal{E})$ as a smooth manifold.

**$\mathfrak{g}, \nabla^{(e)}, \nabla^{(m)}$ on $\mathcal{W}(\mathcal{Y}, \mathcal{E})$ obtained with limiting arguments.**

# Information geometry Markov chains (Nagaoka, 2005)

View $\mathcal{W}(\mathcal{Y}, \mathcal{E})$ as a smooth manifold.

**$\mathfrak{g}, \nabla^{(e)}, \nabla^{(m)}$ on $\mathcal{W}(\mathcal{Y}, \mathcal{E})$ obtained with limiting arguments.**

**Fisher information metric $\mathfrak{g}$**

$$\mathfrak{g}_{ij}(P_\theta) \triangleq \sum_{(y,y') \in \mathcal{E}} Q_\theta(y, y') \partial_i \log P_\theta(y, y') \partial_j \log P_\theta(y, y').$$

**Dual affine connections $\nabla^{(e)}, \nabla^{(m)}$**

$$\Gamma^{(e)}_{ij,k}(P_\theta) \triangleq \sum_{(y,y') \in \mathcal{E}} \partial_i \partial_j \log P_\theta(y, y') \partial_k Q_\theta(y, y'),$$

$$\Gamma^{(m)}_{ij,k}(P_\theta) \triangleq \sum_{(y,y') \in \mathcal{E}} \partial_i \partial_j Q_\theta(y, y') \partial_k \log P_\theta(y, y').$$

# Exponential families of transition kernels (Nagaoka, 2005)

Let $\Theta \subseteq \mathbb{R}^d$, open connected parameter space.

$$\mathcal{V}_e = \left\{ P_\theta \colon \theta = (\theta^1, \ldots, \theta^d) \in \Theta \right\} \subset \mathcal{W}$$

is an e-family with natural parameter $\theta$, whenever there exist functions $K, R_\theta, \psi_\theta, g_1, \ldots, g_d$ such that

$$\log P_\theta(y, y') = K(y, y') + \sum_{i=1}^{d} \theta^i g_i(y, y') + \underbrace{R_\theta(y') - R_\theta(y) - \psi_\theta}_{\text{rescaling terms}}.$$

# Exponential families of transition kernels (Nagaoka, 2005)

Let $\Theta \subseteq \mathbb{R}^d$, open connected parameter space.

$$\mathcal{V}_e = \left\{ P_\theta \colon \theta = (\theta^1, \dots, \theta^d) \in \Theta \right\} \subset \mathcal{W}$$

is an e-family with natural parameter $\theta$, whenever there exist functions $K, R_\theta, \psi_\theta, g_1, \dots, g_d$ such that

$$\log P_\theta(y, y') = K(y, y') + \sum_{i=1}^{d} \theta^i g_i(y, y') + \underbrace{R_\theta(y') - R_\theta(y) - \psi_\theta}_{\text{rescaling terms}}.$$

**Example 1 (Nagaoka, 2005)**

$\mathcal{W}(\mathcal{Y}, \mathcal{E})$ forms an e-family of dimension $|\mathcal{E}| - |\mathcal{Y}|$ . For $\mathcal{E} = \mathcal{Y}^2$,

$$\log P(y, y') = \sum_{i=1}^{|\mathcal{Y}|} \sum_{\substack{j=1 \\ j \neq y_\star}}^{|\mathcal{Y}|} \overbrace{\log \frac{P(i, j) P(j, y_\star)}{P(i, y_\star) P(y_\star, y_\star)}}^{\theta^{ij}} \overbrace{\delta_i(y) \delta_j(y')}^{g_{ij}(y, y')}$$

$$+ \log P(y, y_\star) - \log P(y', y_\star) + \log P(y_\star, y_\star).$$

# Mixture families (Nagaoka, 2005)

We say that $\mathcal{V}_m$ is a mixture family when there exists $C, F_1, \dots, F_d \in \mathcal{F}$, such that $C, C + F_1, \dots, C + F_d$ are affinely independent,

$$\sum_{y,y' \in \mathcal{Y}} C(y, y') = 1, \qquad \sum_{y,y' \in \mathcal{Y}} F_i(y, y') = 0, \forall i \in [d],$$

and

$$\mathcal{V}_m = \left\{ P_\xi \in \mathcal{W} \colon Q_\xi = C + \sum_{i=1}^{d} \xi^i F_i, \xi \in \Xi \right\}$$

where $\Xi = \left\{ \xi \in \mathbb{R}^d \colon Q_\xi(y, y') > 0, \forall (y, y') \in \mathcal{Y}^2 \right\}$, and $Q_\xi$ is the edge measure that pertains to $P_\xi$.

# Mixture families (Nagaoka, 2005)

We say that $\mathcal{V}_m$ is a mixture family when there exists $C, F_1, \dots, F_d \in \mathcal{F}$, such that $C, C + F_1, \dots, C + F_d$ are affinely independent,

$$\sum_{y, y' \in \mathcal{Y}} C(y, y') = 1, \qquad \sum_{y, y' \in \mathcal{Y}} F_i(y, y') = 0, \forall i \in [d],$$

and

$$\mathcal{V}_m = \left\{ P_\xi \in \mathcal{W} \colon Q_\xi = C + \sum_{i=1}^{d} \xi^i F_i, \xi \in \Xi \right\}$$

where $\Xi = \left\{ \xi \in \mathbb{R}^d \colon Q_\xi(y, y') > 0, \forall (y, y') \in \mathcal{Y}^2 \right\}$, and $Q_\xi$ is the edge measure that pertains to $P_\xi$.

**Example 2**

$\mathcal{W}_{\mathrm{rev}}$ (reversible) forms both an m-family and an e-family

(Wolfer and Watanabe, 2021).

# Applications – Inference in Markov chains

Geometric approach has recently lead to finite sample analysis for:

1. Parameter estimation problem in Markov chains in HMMs Hayashi and Watanabe (2016); Hayashi (2022).

# Applications – Inference in Markov chains

Geometric approach has recently lead to finite sample analysis for:

1. Parameter estimation problem in Markov chains in HMMs
   Hayashi and Watanabe (2016); Hayashi (2022).
2. Hypothesis testing problem
   Watanabe and Hayashi (2017).

# Applications – Inference in Markov chains

Geometric approach has recently lead to finite sample analysis for:

1. Parameter estimation problem in Markov chains in HMMs
   Hayashi and Watanabe (2016); Hayashi (2022).
2. Hypothesis testing problem
   Watanabe and Hayashi (2017).
3. Local equivalence problem in hidden Markov models
   Hayashi (2019).

# Applications – Inference in Markov chains

Geometric approach has recently lead to finite sample analysis for:

1. Parameter estimation problem in Markov chains in HMMs
   Hayashi and Watanabe (2016); Hayashi (2022).
2. Hypothesis testing problem
   Watanabe and Hayashi (2017).
3. Local equivalence problem in hidden Markov models
   Hayashi (2019).
4. Chernoff and Hoeffding bounds with improved pre-factor
   Moulos and Anantharam (2019).

# Data-processing & Lumping

# Data-processing – Distribution setting

$$Y_1, Y_2, \ldots, Y_n \sim \mu^{\otimes n}$$

# Data-processing – Distribution setting

$$Y_1, Y_2, \ldots, Y_n \sim \mu^{\otimes n}$$

**Possibly randomized function $\phi$**

$$\phi \colon \mathcal{Y} \to \mathcal{X}$$

# Data-processing – Distribution setting

$$Y_1, Y_2, \ldots, Y_n \sim \mu^{\otimes n}$$

**Possibly randomized function $\phi$**

$$\phi \colon \mathcal{Y} \to \mathcal{X}$$

**Operational definition of data-processing**

$$\phi(Y_1), \phi(Y_2), \ldots, \phi(Y_n) \sim (\phi_\star \mu)^{\otimes n}$$

# Data-processing – Distribution setting

$$Y_1, Y_2, \ldots, Y_n \sim \mu^{\otimes n}$$

**Possibly randomized function $\phi$**

$$\phi \colon \mathcal{Y} \to \mathcal{X}$$

**Operational definition of data-processing**

$$\phi(Y_1), \phi(Y_2), \ldots, \phi(Y_n) \sim (\phi_\star \mu)^{\otimes n}$$

**Memoryless channel definition**

$$\phi_\star \cong W \colon \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{X}), \qquad \phi_\star \mu = \mu W$$

# Data-processing – Distribution setting

$$Y_1, Y_2, \dots, Y_n \sim \mu^{\otimes n}$$

**Possibly randomized function $\phi$**

$$\phi \colon \mathcal{Y} \to \mathcal{X}$$

**Operational definition of data-processing**

$$\phi(Y_1), \phi(Y_2), \dots, \phi(Y_n) \sim (\phi_\star \mu)^{\otimes n}$$

**Memoryless channel definition**

$$\phi_\star \cong W \colon \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{X}), \qquad \phi_\star \mu = \mu W$$

**Monotonicity of information**

$$D(\mu \| \nu) \geq D(\mu W \| \nu W)$$

# Data-processing – Distribution setting

$$Y_1, Y_2, \ldots, Y_n \sim \mu^{\otimes n}$$

**Possibly randomized function $\phi$**

$$\phi\colon \mathcal{Y} \to \mathcal{X}$$

**Operational definition of data-processing**

$$\phi(Y_1), \phi(Y_2), \ldots, \phi(Y_n) \sim (\phi_\star \mu)^{\otimes n}$$

**Memoryless channel definition**

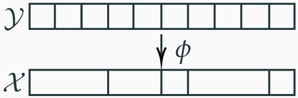
$$\phi_\star \cong W\colon \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{X}), \qquad \phi_\star\mu = \mu W$$

**Monotonicity of information**

$$D\left(\mu \| \nu\right) \geq D\left(\mu W \| \nu W\right)$$

**Lumping**

When $\phi$ is deterministic and $|\mathcal{X}| \leq |\mathcal{Y}|$.



# Lumpability of Markov chains

**Markovian setting**

$$Y_1, Y_2, \ldots, Y_n \sim P \in \mathcal{W}(\mathcal{Y}, \mathcal{E})$$

# Lumpability of Markov chains

**Markovian setting**

$$Y_1, Y_2, \ldots, Y_n \sim P \in \mathcal{W}(\mathcal{Y}, \mathcal{E})$$

**Lumping function $\kappa$**

$$\kappa \colon \mathcal{Y} \to \mathcal{X}, \biguplus_{x \in \mathcal{X}} \mathcal{S}_x = \mathcal{Y}$$

# Lumpability of Markov chains

**Markovian setting**

$$Y_1, Y_2, \ldots, Y_n \sim P \in \mathcal{W}(\mathcal{Y}, \mathcal{E})$$

**Lumping function** $\kappa$

$$\kappa \colon \mathcal{Y} \to \mathcal{X}, \biguplus_{x \in \mathcal{X}} \mathcal{S}_x = \mathcal{Y}$$

**Operational definition of data-processing**

$$\kappa(Y_1), \kappa(Y_2), \ldots, \kappa(Y_n) \sim \kappa_\star P \ ?$$

# Lumpability of Markov chains

**Markovian setting**

$$Y_1, Y_2, \ldots, Y_n \sim P \in \mathcal{W}(\mathcal{Y}, \mathcal{E})$$

**Lumping function $\kappa$**

$$\kappa \colon \mathcal{Y} \to \mathcal{X}, \biguplus_{x \in \mathcal{X}} \mathcal{S}_x = \mathcal{Y}$$

**Operational definition of data-processing**

$$\kappa(Y_1), \kappa(Y_2), \ldots, \kappa(Y_n) \sim \kappa_\star P \ \ ?$$

When $(\kappa(Y_t))_{t \in \mathbb{N}}$ is Markovian, we say $P$ is lumpable: $P \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$.

# Lumpability characterization & example

**Characterization (Kemeny and Snell, 1983, Theorem 6.3.2)**

$P \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$ iff for all $x, x' \in \mathcal{X}$, and for all $y_1, y_2 \in \mathcal{S}_x$,

$$P(y_1, \mathcal{S}_{x'}) = P(y_2, \mathcal{S}_{x'}) \triangleq \kappa_\star P(x, x').$$

# Lumpability characterization & example

**Characterization (Kemeny and Snell, 1983, Theorem 6.3.2)**

$P \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$ iff for all $x, x' \in \mathcal{X}$, and for all $y_1, y_2 \in \mathcal{S}_x$,

$$P(y_1, \mathcal{S}_{x'}) = P(y_2, \mathcal{S}_{x'}) \triangleq \kappa_\star P(x, x').$$

**Example 3**

$\mathcal{Y} = \{0, 1, 2\}, \mathcal{X} = \{0, 1\}$

$\kappa \colon \mathcal{Y} \to \mathcal{X}, \qquad \kappa(0) = 0, \kappa(1) = \kappa(2) = 1.$

# Lumpability characterization & example

**Characterization (Kemeny and Snell, 1983, Theorem 6.3.2)**

$P \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$ iff for all $x, x' \in \mathcal{X}$, and for all $y_1, y_2 \in \mathcal{S}_x$,

$$P(y_1, \mathcal{S}_{x'}) = P(y_2, \mathcal{S}_{x'}) \triangleq \kappa_\star P(x, x').$$

**Example 3**

$\mathcal{Y} = \{0, 1, 2\}$, $\mathcal{X} = \{0, 1\}$

$\kappa \colon \mathcal{Y} \to \mathcal{X}, \qquad \kappa(0) = 0, \kappa(1) = \kappa(2) = 1.$

$$P = \left(\begin{array}{c|cc} 4/5 & 1/10 & 1/10 \\ \hline 1/2 & 3/10 & 1/5 \\ 1/2 & 1/5 & 3/10 \end{array}\right),$$

# Lumpability characterization & example

**Characterization (Kemeny and Snell, 1983, Theorem 6.3.2)**

$P \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$ iff for all $x, x' \in \mathcal{X}$, and for all $y_1, y_2 \in \mathcal{S}_x$,

$$P(y_1, \mathcal{S}_{x'}) = P(y_2, \mathcal{S}_{x'}) \triangleq \kappa_\star P(x, x').$$

**Example 3**

$\mathcal{Y} = \{0, 1, 2\}, \mathcal{X} = \{0, 1\}$

$\kappa \colon \mathcal{Y} \to \mathcal{X}, \qquad \kappa(0) = 0, \kappa(1) = \kappa(2) = 1.$

$$P = \begin{pmatrix} 4/5 & 1/10 & 1/10 \\ 1/2 & 3/10 & 1/5 \\ 1/2 & 1/5 & 3/10 \end{pmatrix}, \qquad \kappa_\star P = \begin{pmatrix} 4/5 & 1/5 \\ 1/2 & 1/2 \end{pmatrix}.$$

# Markov Embeddings

# Markov morphisms – Distribution setting

**Definition 4 (Čencov (1978); Campbell (1986))**

Let the partition $\biguplus_{x \in \mathcal{X}} \mathcal{S}_x = \mathcal{Y}$.

To each $x \in \mathcal{X}$, we associate $W^x \in \mathcal{P}(\mathcal{Y})$ concentrated on $\mathcal{S}_x$.

$$W_\star \colon \mathcal{P}_+(\mathcal{X}) \to \mathcal{P}_+(\mathcal{Y}), \mu \mapsto W_\star \mu(y) = \sum_{x \in \mathcal{X}} W^x(y) \mu(x), \forall y \in \mathcal{Y}.$$

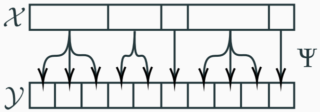
# Markov morphisms – Distribution setting

**Definition 4 (Čencov (1978); Campbell (1986))**

Let the partition $\biguplus_{x \in \mathcal{X}} \mathcal{S}_x = \mathcal{Y}$.

To each $x \in \mathcal{X}$, we associate $W^x \in \mathcal{P}(\mathcal{Y})$ concentrated on $\mathcal{S}_x$.

$$W_\star \colon \mathcal{P}_+(\mathcal{X}) \to \mathcal{P}_+(\mathcal{Y}), \mu \mapsto W_\star\mu(y) = \sum_{x \in \mathcal{X}} W^x(y)\mu(x), \forall y \in \mathcal{Y}.$$

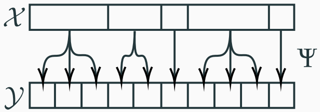


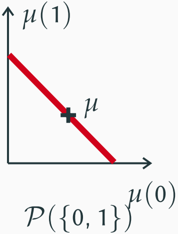
**Definition 4 (Čencov (1978); Campbell (1986))**

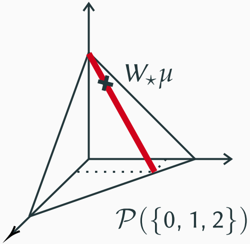
Let the partition $\biguplus_{x \in \mathcal{X}} \mathcal{S}_x = \mathcal{Y}$.

To each $x \in \mathcal{X}$, we associate $W^x \in \mathcal{P}(\mathcal{Y})$ concentrated on $\mathcal{S}_x$.

$$W_\star \colon \mathcal{P}_+(\mathcal{X}) \to \mathcal{P}_+(\mathcal{Y}), \mu \mapsto W_\star\mu(y) = \sum_{x \in \mathcal{X}} W^x(y)\mu(x), \forall y \in \mathcal{Y}.$$



# Congruent linear embeddings

**Definition 5 (Congruent linear embeddings)**
For a statistic $\kappa\colon \mathcal{Y} \to \mathcal{X}$, a $\kappa$-congruent embedding $K_\star$ is a map $K_\star\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{Y}}$ that verifies:

# Congruent linear embeddings

**Definition 5 (Congruent linear embeddings)**
For a statistic $\kappa\colon \mathcal{Y} \to \mathcal{X}$, a $\kappa$-congruent embedding $K_\star$ is a map $K_\star\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{Y}}$ that verifies:

$(i)$ $K_\star$ is linear.

# Congruent linear embeddings

**Definition 5 (Congruent linear embeddings)**
For a statistic $\kappa\colon \mathcal{Y} \to \mathcal{X}$, a $\kappa$-congruent embedding $K_\star$ is a map $K_\star\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{Y}}$ that verifies:

(*i*) $K_\star$ is linear.

(*ii*) $K_\star$ is monotonic (preserves non-negativity of measure).

# Congruent linear embeddings

**Definition 5 (Congruent linear embeddings)**
For a statistic $\kappa\colon \mathcal{Y} \to \mathcal{X}$, a $\kappa$-congruent embedding $K_\star$ is a map $K_\star\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{Y}}$ that verifies:

(*i*) $K_\star$ is linear.

(*ii*) $K_\star$ is monotonic (preserves non-negativity of measure).

(*iii*) $K_\star$ is a right inverse of $\kappa_\star$.

**Definition 5 (Congruent linear embeddings)**
For a statistic $\kappa\colon \mathcal{Y} \to \mathcal{X}$, a $\kappa$-congruent embedding $K_\star$ is a map $K_\star\colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{Y}}$ that verifies:

- $(i)$ $K_\star$ is linear.
- $(ii)$ $K_\star$ is monotonic (preserves non-negativity of measure).
- $(iii)$ $K_\star$ is a right inverse of $\kappa_\star$.

**Property (Ay et al., 2017, Example 5.2)**

Congruent embedding iff (congruent) Markov morphism.

# Extension: Markov embeddings of Markov kernels

Let $\kappa \colon \mathcal{Y} \to \mathcal{X}$ inducing the partition $\biguplus_{x \in \mathcal{X}} \mathcal{S}_x = \mathcal{Y}$.

# Extension: Markov embeddings of Markov kernels

Let $\kappa \colon \mathcal{Y} \to \mathcal{X}$ inducing the partition $\biguplus_{x \in \mathcal{X}} \mathcal{S}_x = \mathcal{Y}$.

**Definition 6 ($\kappa$-compatible Markov embedding)**

$$\Lambda_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \supset \mathcal{V} \to \mathcal{W}(\mathcal{Y}, \mathcal{E})$$
$$P \mapsto P(\kappa(y), \kappa(y'))\Lambda(y, y'), \forall (y, y') \in \mathcal{E},$$

# Extension: Markov embeddings of Markov kernels

Let $\kappa \colon \mathcal{Y} \to \mathcal{X}$ inducing the partition $\biguplus_{x \in \mathcal{X}} \mathcal{S}_x = \mathcal{Y}$.

**Definition 6 ($\kappa$-compatible Markov embedding)**

$$\begin{aligned} \Lambda_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \supset \mathcal{V} &\to \mathcal{W}(\mathcal{Y}, \mathcal{E}) \\ P &\mapsto P(\kappa(y), \kappa(y'))\Lambda(y, y'), \forall (y, y') \in \mathcal{E}, \end{aligned}$$

(*i*) $\kappa$ and $\mathcal{E}$ satisfy $\kappa_2(\mathcal{E}) = \mathcal{D}$ and $\mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) \neq \emptyset$.

(*ii*) $\Lambda \in \mathcal{F}_+(\mathcal{Y}, \mathcal{E})$.

(*iii*) $\forall y \in \mathcal{Y}, x' \in \mathcal{X}, (\kappa(y), x') \in \mathcal{D} \implies (\Lambda(y, y'))_{y' \in \mathcal{S}_{x'}} \in \mathcal{P}(\mathcal{S}_{x'})$.

# Extension: Markov embeddings of Markov kernels

Let $\kappa \colon \mathcal{Y} \to \mathcal{X}$ inducing the partition $\biguplus_{x \in \mathcal{X}} \mathcal{S}_x = \mathcal{Y}$.

**Definition 6 ($\kappa$-compatible Markov embedding)**

$$\begin{aligned}\Lambda_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \supset \mathcal{V} &\to \mathcal{W}(\mathcal{Y}, \mathcal{E}) \\ P &\mapsto P(\kappa(y), \kappa(y'))\Lambda(y, y'), \forall (y, y') \in \mathcal{E},\end{aligned}$$

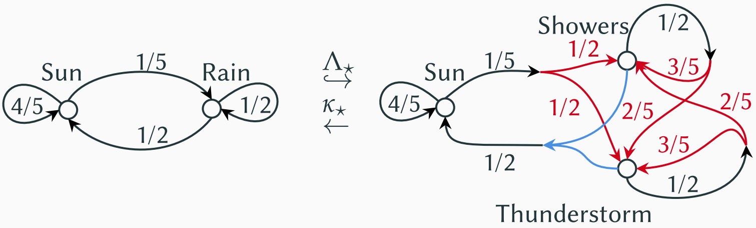
(*i*) $\kappa$ and $\mathcal{E}$ satisfy $\kappa_2(\mathcal{E}) = \mathcal{D}$ and $\mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) \neq \emptyset$.

(*ii*) $\Lambda \in \mathcal{F}_+(\mathcal{Y}, \mathcal{E})$.

(*iii*) $\forall y \in \mathcal{Y}, x' \in \mathcal{X}, (\kappa(y), x') \in \mathcal{D} \implies (\Lambda(y, y'))_{y' \in \mathcal{S}_{x'}} \in \mathcal{P}(\mathcal{S}_{x'})$.

$$\Lambda = \begin{pmatrix} W_{1,1} & W_{1,2} & \cdots & & W_{1,n} \\ \vdots & & & & \vdots \\ W_{x,1} & \cdots & W_{x,x'} & \cdots & W_{x,n} \\ \vdots & & & & \vdots \\ W_{n,1} & & \cdots & & W_{n,n} \end{pmatrix}.$$

# Example: weather model




$$P = \begin{pmatrix} 4/5 & 1/5 \\ 1/2 & 1/2 \end{pmatrix}, \Lambda = \left(\begin{array}{c:cc} 1 & 1/2 & 1/2 \\ \hdashline 1 & 3/5 & 2/5 \\ 1 & 2/5 & 3/5 \end{array}\right),$$

$$\Lambda_\star P = \left(\begin{array}{c:cc} 4/5 & 1/10 & 1/10 \\ \hdashline 1/2 & 3/10 & 1/5 \\ 1/2 & 1/5 & 3/10 \end{array}\right).$$

# Properties of Markov embeddings

**Related work on conditional models**
Lebanon (2004, 2005); Montúfar, Rauh, and Ay (2014).

# Properties of Markov embeddings

**Related work on conditional models**
Lebanon (2004, 2005); Montúfar, Rauh, and Ay (2014).

**Lemma 7 (Operational definition)**
*Embedded trajectory can be simulated from the original trajectory.*

# Properties of Markov embeddings

**Related work on conditional models**
Lebanon (2004, 2005); Montúfar, Rauh, and Ay (2014).

**Lemma 7 (Operational definition)**
*Embedded trajectory can be simulated from the original trajectory.*

$$X_1 \qquad \dots \qquad X_k \qquad\qquad\qquad \sim P$$

# Properties of Markov embeddings

**Related work on conditional models**
Lebanon (2004, 2005); Montúfar, Rauh, and Ay (2014).

**Lemma 7 (Operational definition)**
*Embedded trajectory can be simulated from the original trajectory.*

$$
\begin{array}{cccc}
X_1 & \dots & X_k & \sim P \\
\downarrow & & \downarrow & \Psi_\Lambda \text{ randomized mapping}
\end{array}
$$

# Properties of Markov embeddings

**Related work on conditional models**
Lebanon (2004, 2005); Montúfar, Rauh, and Ay (2014).

**Lemma 7 (Operational definition)**
*Embedded trajectory can be simulated from the original trajectory.*

$$\begin{array}{ccccc} X_1 & \dots & X_k & & \sim P \\ \downarrow & & \downarrow & & \Psi_\Lambda \text{ randomized mapping} \\ \Psi_\Lambda(X_1) & \dots & \Psi_\Lambda(X_k) & & \sim \Lambda_\star P \end{array}$$

# Properties of Markov embeddings

**Related work on conditional models**
Lebanon (2004, 2005); Montúfar, Rauh, and Ay (2014).

**Lemma 7 (Operational definition)**
*Embedded trajectory can be simulated from the original trajectory.*

$$
\begin{array}{ccccl}
X_1 & \dots & X_k & & \sim P \\
\downarrow & & \downarrow & & \Psi_\Lambda \text{ randomized mapping} \\
\Psi_\Lambda(X_1) & \dots & \Psi_\Lambda(X_k) & & \sim \Lambda_\star P
\end{array}
$$

**Lemma 8 (Geometry preservation)**
*$\bar{P}_\theta, \bar{P}_{\theta'} \in \mathcal{V}$ and $\Lambda_\star \colon \mathcal{V} \to \mathcal{W}(\mathcal{Y}, \mathcal{E})$ a Markov embedding, $P_\theta \triangleq \Lambda_\star \bar{P}_\theta$ and $P_{\theta'} \triangleq \Lambda_\star \bar{P}_{\theta'}$.*

$$
\mathfrak{g}_{ij}(P_\theta) = \mathfrak{g}_{ij}(\bar{P}_\theta), \qquad D\left(P_\theta \| P_{\theta'}\right) = D\left(\bar{P}_\theta \| \bar{P}_{\theta'}\right).
$$

$$
\Gamma^{(e)}_{ij,k}(P_\theta) = \Gamma^{(e)}_{ij,k}(\bar{P}_\theta), \qquad \Gamma^{(m)}_{ij,k}(P_\theta) = \Gamma^{(m)}_{ij,k}(\bar{P}_\theta).
$$

# Properties of Markov embeddings

**Related work on conditional models**
Lebanon (2004, 2005); Montúfar, Rauh, and Ay (2014).

**Lemma 7 (Operational definition)**
*Embedded trajectory can be simulated from the original trajectory.*

$$\begin{array}{ccccc} X_1 & \dots & X_k & & \sim P \\ \downarrow & & \downarrow & \Psi_\Lambda \text{ randomized mapping} & \\ \Psi_\Lambda(X_1) & \dots & \Psi_\Lambda(X_k) & & \sim \Lambda_\star P \end{array}$$

**Lemma 8 (Geometry preservation)**
*$\bar{P}_\theta, \bar{P}_{\theta'} \in \mathcal{V}$ and $\Lambda_\star \colon \mathcal{V} \to \mathcal{W}(\mathcal{Y}, \mathcal{E})$ a Markov embedding, $P_\theta \triangleq \Lambda_\star \bar{P}_\theta$ and $P_{\theta'} \triangleq \Lambda_\star \bar{P}_{\theta'}$.*

$$\mathfrak{g}_{ij}(P_\theta) = \mathfrak{g}_{ij}(\bar{P}_\theta), \qquad D\left(P_\theta \| P_{\theta'}\right) = D\left(\bar{P}_\theta \| \bar{P}_{\theta'}\right).$$
$$\Gamma^{(e)}_{ij,k}(P_\theta) = \Gamma^{(e)}_{ij,k}(\bar{P}_\theta), \qquad \Gamma^{(m)}_{ij,k}(P_\theta) = \Gamma^{(m)}_{ij,k}(\bar{P}_\theta).$$

*Furthermore, $\Lambda_\star$ is e-geodesic affine.*

# Properties of Markov embeddings

**Related work on conditional models**
Lebanon (2004, 2005); Montúfar, Rauh, and Ay (2014).

**Lemma 7 (Operational definition)**
*Embedded trajectory can be simulated from the original trajectory.*

$$\begin{array}{cccc} X_1 & \dots & X_k & \sim P \\ \downarrow & & \downarrow & \Psi_\Lambda \text{ randomized mapping} \\ \Psi_\Lambda(X_1) & \dots & \Psi_\Lambda(X_k) & \sim \Lambda_\star P \end{array}$$

**Lemma 8 (Geometry preservation)**
$\bar{P}_\theta, \bar{P}_{\theta'} \in \mathcal{V}$ *and* $\Lambda_\star \colon \mathcal{V} \to \mathcal{W}(\mathcal{Y}, \mathcal{E})$ *a Markov embedding,* $P_\theta \triangleq \Lambda_\star \bar{P}_\theta$ *and* $P_{\theta'} \triangleq \Lambda_\star \bar{P}_{\theta'}$.

$$\mathfrak{g}_{ij}(P_\theta) = \mathfrak{g}_{ij}(\bar{P}_\theta), \qquad D\left(P_\theta \| P_{\theta'}\right) = D\left(\bar{P}_\theta \| \bar{P}_{\theta'}\right).$$

$$\Gamma^{(e)}_{ij,k}(P_\theta) = \Gamma^{(e)}_{ij,k}(\bar{P}_\theta), \qquad \Gamma^{(m)}_{ij,k}(P_\theta) = \Gamma^{(m)}_{ij,k}(\bar{P}_\theta).$$

*Furthermore,* $\Lambda_\star$ *is e-geodesic affine.*

**Important observation**
Unlike distribution setting, Markov embeddings are not m-geodesic affine.

# Extension: Congruent linear embeddings of Markov kernels

Introduce vector space of lumpable matrices $\mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E})$.

# Extension: Congruent linear embeddings of Markov kernels

Introduce vector space of lumpable matrices $\mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E})$.

**Definition 9 ($\kappa$-congruent embedding)**

$$K_\star \colon \mathcal{F}(\mathcal{X} = \kappa(\mathcal{Y}), \mathcal{D} = \kappa_2(\mathcal{E})) \to \mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E}).$$

# Extension: Congruent linear embeddings of Markov kernels

Introduce vector space of lumpable matrices $\mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E})$.

**Definition 9 ($\kappa$-congruent embedding)**

$$K_\star \colon \mathcal{F}(\mathcal{X} = \kappa(\mathcal{Y}), \mathcal{D} = \kappa_2(\mathcal{E})) \to \mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E}).$$

$(i)$ $K_\star$ is a linear map.

# Extension: Congruent linear embeddings of Markov kernels

Introduce vector space of lumpable matrices $\mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E})$.

**Definition 9 ($\kappa$-congruent embedding)**

$$K_\star \colon \mathcal{F}(\mathcal{X} = \kappa(\mathcal{Y}), \mathcal{D} = \kappa_2(\mathcal{E})) \to \mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E}).$$

$(i)$ $K_\star$ is a linear map.

$(ii)$ $K_\star$ is monotonic,

$$A \in \mathcal{F}(\mathcal{X}, \mathcal{D}), A \geq 0 \implies K_\star A \geq 0.$$

# Extension: Congruent linear embeddings of Markov kernels

Introduce vector space of lumpable matrices $\mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E})$.

**Definition 9 ($\kappa$-congruent embedding)**

$$K_\star \colon \mathcal{F}(\mathcal{X} = \kappa(\mathcal{Y}), \mathcal{D} = \kappa_2(\mathcal{E})) \to \mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E}).$$

$(i)$ $K_\star$ is a linear map.

$(ii)$ $K_\star$ is monotonic,

$$A \in \mathcal{F}(\mathcal{X}, \mathcal{D}), A \geq 0 \implies K_\star A \geq 0.$$

$(iii)$ $K_\star$ preserves irreducibility,

$$A \in \mathcal{F}_+(\mathcal{X}, \mathcal{D}) \implies K_\star A \in \mathcal{F}_+(\mathcal{Y}, \mathcal{E}).$$

# Extension: Congruent linear embeddings of Markov kernels

Introduce vector space of lumpable matrices $\mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E})$.

**Definition 9 ($\kappa$-congruent embedding)**

$$K_\star \colon \mathcal{F}(\mathcal{X} = \kappa(\mathcal{Y}), \mathcal{D} = \kappa_2(\mathcal{E})) \to \mathcal{F}_\kappa(\mathcal{Y}, \mathcal{E}).$$

$(i)$ $K_\star$ is a linear map.

$(ii)$ $K_\star$ is monotonic,

$$A \in \mathcal{F}(\mathcal{X}, \mathcal{D}), A \geq 0 \implies K_\star A \geq 0.$$

$(iii)$ $K_\star$ preserves irreducibility,

$$A \in \mathcal{F}_+(\mathcal{X}, \mathcal{D}) \implies K_\star A \in \mathcal{F}_+(\mathcal{Y}, \mathcal{E}).$$

$(iv)$ $K_\star$ is a right inverse of $\kappa_\star$, i.e. for any $\forall A \in \mathcal{F}(\mathcal{X}, \mathcal{D})$,

$$\kappa_\star K_\star A = A.$$

# Congruent Embeddings are Compatible Markov Embeddings

**Theorem 10**
*Let* $(\mathcal{X}, \mathcal{D})$, $(\mathcal{Y}, \mathcal{E})$ *be strongly connected digraphs, and* $\kappa \colon \mathcal{Y} \to \mathcal{X}$ *a lumping function, such that* $\kappa_2(\mathcal{E}) = \mathcal{D}$, *and* $\mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) \neq \emptyset$.

# Congruent Embeddings are Compatible Markov Embeddings

**Theorem 10**
*Let $(\mathcal{X}, \mathcal{D})$, $(\mathcal{Y}, \mathcal{E})$ be strongly connected digraphs, and $\kappa \colon \mathcal{Y} \to \mathcal{X}$ a lumping function, such that $\kappa_2(\mathcal{E}) = \mathcal{D}$, and $\mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) \neq \emptyset$. Then*

*$K_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \to \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$ is a $\kappa$-congruent embedding,*

*if and only if*

*$K_\star$ is a $\kappa$-compatible Markov embedding.*

# Example: Hudson expansions

# Hudson expansions

Natural expansions, inverse of lumping, considered in Kemeny and Snell (1983, Section 6.5,p.140).

$$H_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \to \mathcal{W}_h(\mathcal{D}, H_{\mathcal{D}}).$$

When

$$X_1, X_2, \dots, X_t, \dots \sim \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D}),$$

then sliding window observations

$$(X_1, X_2), (X_2, X_3), \dots, (X_t, X_{t+1}), \dots$$

also forms a Markov chain with kernel $P = H_\star \bar{P}$ and $H_\star \bar{\pi}(e) = \bar{Q}(e)$.

# Hudson expansions

Natural expansions, inverse of lumping, considered in Kemeny and Snell (1983, Section 6.5,p.140).

$$H_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \to \mathcal{W}_h(\mathcal{D}, H_\mathcal{D}).$$

When

$$X_1, X_2, \dots, X_t, \dots \sim \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D}),$$

then sliding window observations

$$(X_1, X_2), (X_2, X_3), \dots, (X_t, X_{t+1}), \dots$$

also forms a Markov chain with kernel $P = H_\star \bar{P}$ and $H_\star \bar{\pi}(e) = \bar{Q}(e)$.

**Theorem 11**

*(i) $H_\star$ is a Markov embedding.*

# Hudson expansions

Natural expansions, inverse of lumping, considered in Kemeny and Snell (1983, Section 6.5,p.140).

$$H_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \to \mathcal{W}_h(\mathcal{D}, H_\mathcal{D}).$$

When

$$X_1, X_2, \dots, X_t, \dots \sim \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D}),$$

then sliding window observations

$$(X_1, X_2), (X_2, X_3), \dots, (X_t, X_{t+1}), \dots$$

also forms a Markov chain with kernel $P = H_\star \bar{P}$ and $H_\star \bar{\pi}(e) = \bar{Q}(e)$.

**Theorem 11**

*(i) $H_\star$ is a Markov embedding.*

*(ii) $H_\star$ is not m-geodesic affine.*

# Hudson expansions

Natural expansions, inverse of lumping, considered in Kemeny and Snell (1983, Section 6.5,p.140).

$$H_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \to \mathcal{W}_h(\mathcal{D}, H_\mathcal{D}).$$

When

$$X_1, X_2, \dots, X_t, \dots \sim \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D}),$$

then sliding window observations

$$(X_1, X_2), (X_2, X_3), \dots, (X_t, X_{t+1}), \dots$$

also forms a Markov chain with kernel $P = H_\star \bar{P}$ and $H_\star \bar{\pi}(e) = \bar{Q}(e)$.

**Theorem 11**

(i) *$H_\star$ is a Markov embedding.*

(ii) *$H_\star$ is not m-geodesic affine.*

(iii) *Can view $H_\star \mathcal{W}(\mathcal{X}, \mathcal{D})$ as $1^{st}$-order sub-family of $2^{nd}$-order kernels.*

# Hudson expansions

Natural expansions, inverse of lumping, considered in Kemeny and Snell (1983, Section 6.5,p.140).

$$H_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \to \mathcal{W}_h(\mathcal{D}, H_{\mathcal{D}}).$$

When

$$X_1, X_2, \dots, X_t, \dots \sim \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D}),$$

then sliding window observations

$$(X_1, X_2), (X_2, X_3), \dots, (X_t, X_{t+1}), \dots$$

also forms a Markov chain with kernel $P = H_\star \bar{P}$ and $H_\star \bar{\pi}(e) = \bar{Q}(e)$.

**Theorem 11**

(i) *$H_\star$ is a Markov embedding.*

(ii) *$H_\star$ is not m-geodesic affine.*

(iii) *Can view $H_\star \mathcal{W}(\mathcal{X}, \mathcal{D})$ as 1st-order sub-family of 2nd-order kernels.*

(iv) *Theory extends to higher-order.*

# Information Geometry of Lumpable Kernels

# Linear family of kernels that lump into prescribed $\bar{P}_0$

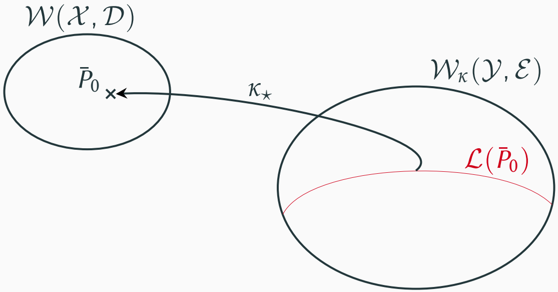
**Observation**
$\mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$ is generally not e-family or m-family.

# Linear family of kernels that lump into prescribed $\bar{P}_0$

**Observation**
$\mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$ is generally not e-family or m-family.

Let $\bar{P}_0 \in \mathcal{W}(\mathcal{X}, \mathcal{D})$, and

$$\mathcal{L}(\bar{P}_0) \triangleq \{P \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) \colon \kappa_\star P = \bar{P}_0\},$$



# Linear family of kernels that lump into prescribed $\bar{P}_0$

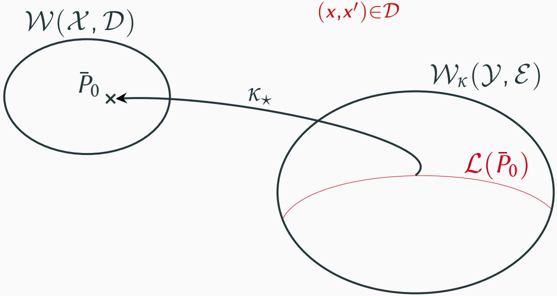
**Observation**
$\mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$ is generally not e-family or m-family.

Let $\bar{P}_0 \in \mathcal{W}(\mathcal{X}, \mathcal{D})$, and

$$\mathcal{L}(\bar{P}_0) \triangleq \{P \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) \colon \kappa_\star P = \bar{P}_0\},$$

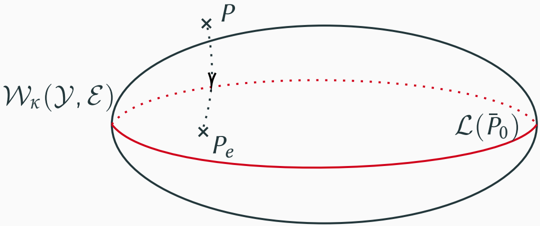
**Lemma 12**
*$\mathcal{L}(\bar{P}_0)$ forms an m-family in $\mathcal{W}(\mathcal{Y}, \mathcal{E})$, with*

$$\dim \mathcal{L}(\bar{P}_0) = |\mathcal{E}| - \sum_{(x,x') \in \mathcal{D}} |\mathcal{S}_x|.$$



# Application: Maximum Entropy Principle

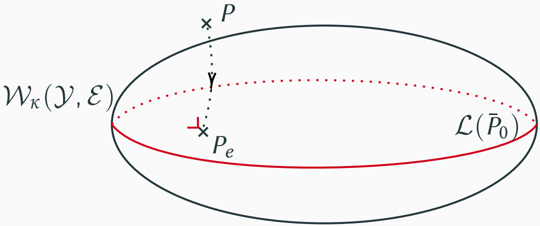
$$P_e \triangleq \operatorname*{arg\,min}_{P' \in \mathcal{L}(\bar{P}_0)} D\left(P' \| P\right).$$



# Application: Maximum Entropy Principle

$$P_e \triangleq \underset{P' \in \mathcal{L}(\bar{P}_0)}{\arg\min}\, D\left(P' \| P\right).$$

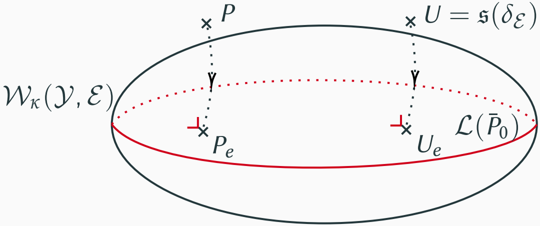
e-geodesic going through $P$ and $P_e$ is $\perp$ to $\mathcal{L}(\bar{P}_0)$.



# Application: Maximum Entropy Principle

$$P_e \triangleq \underset{P' \in \mathcal{L}(\bar{P}_0)}{\arg\min}\, D\left(P' \| P\right).$$

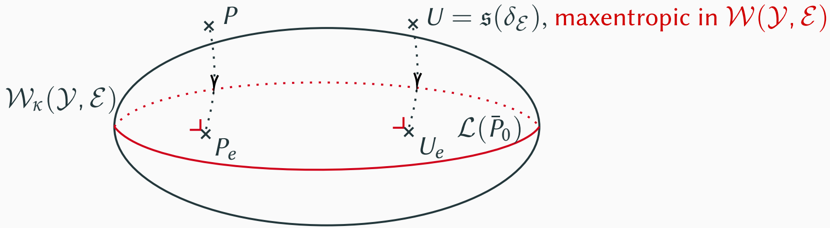
e-geodesic going through $P$ and $P_e$ is $\perp$ to $\mathcal{L}(\bar{P}_0)$.



# Application: Maximum Entropy Principle

$$P_e \triangleq \arg\min_{P' \in \mathcal{L}(\bar{P}_0)} D\left(P' \| P\right).$$

e-geodesic going through $P$ and $P_e$ is $\perp$ to $\mathcal{L}(\bar{P}_0)$.



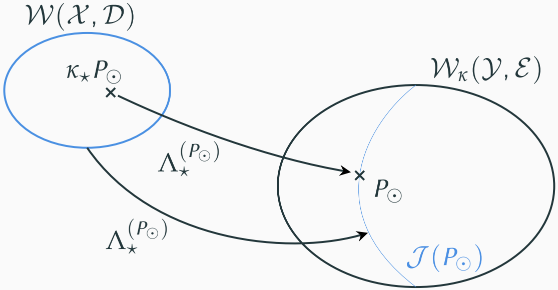

Entropy rate: $H(P) \triangleq \lim_{k \to \infty} \frac{1}{k} H(Y_1, Y_2, \dots, Y_k)$

and rewrite

$$U_e = \arg\min_{P' \in \mathcal{L}(\bar{P}_0)} \left\{ -H(P') - \overbrace{\mathbb{E}_{(Y,Y') \sim Q'} \left[ \log U(Y, Y') \right]}^{-\log \rho(\mathfrak{s}(\delta_{\mathcal{E}}))} \right\} = \arg\max_{P' \in \mathcal{L}(\bar{P}_0)} H(P').$$

# e-family of embedded kernels at some prescribed origin $P_\odot$

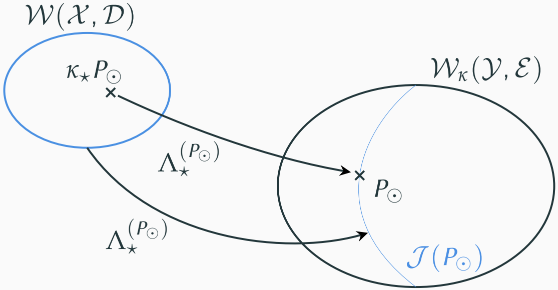
$P_\odot \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$, and $\Lambda_\star^{(P_\odot)}$ canonical emb. (e.g. $\Lambda_\star^{(P_\odot)} \kappa_\star P_\odot = P_\odot$).



# e-family of embedded kernels at some prescribed origin $P_\odot$

$P_\odot \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$, and $\Lambda_\star^{(P_\odot)}$ canonical emb. (e.g. $\Lambda_\star^{(P_\odot)} \kappa_\star P_\odot = P_\odot$).

$$\mathcal{J}(P_\odot) \triangleq \left\{ \Lambda_\star^{(P_\odot)} \bar{P} \colon \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D}) \right\} \subset \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}).$$



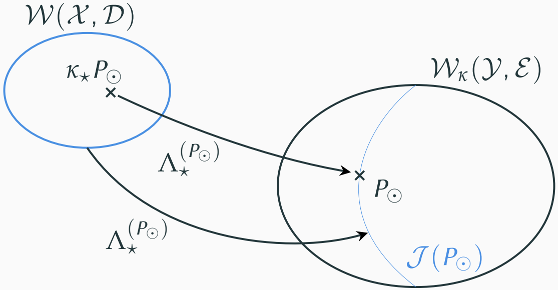
# e-family of embedded kernels at some prescribed origin $P_\odot$

$P_\odot \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$, and $\Lambda_\star^{(P_\odot)}$ canonical emb. (e.g. $\Lambda_\star^{(P_\odot)} \kappa_\star P_\odot = P_\odot$).

$$\mathcal{J}(P_\odot) \triangleq \left\{ \Lambda_\star^{(P_\odot)} \bar{P} \colon \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D}) \right\} \subset \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}).$$

**Lemma 13**
*$\mathcal{J}(P_\odot)$ forms an e-family in $\mathcal{W}(\mathcal{Y}, \mathcal{E})$, with*

$$\dim \mathcal{J}(P_\odot) = |\mathcal{D}| - |\mathcal{X}|.$$



# Foliated manifold of lumpable kernels

**Theorem 14**

*For any fixed* $\bar{P}_0 \in \mathcal{W}(\mathcal{X}, \mathcal{D})$,

$$\mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) = \biguplus_{P_\odot \in \mathcal{L}(\bar{P}_0)} \mathcal{J}(P_\odot).$$

# Foliated manifold of lumpable kernels

**Theorem 14**

*For any fixed* $\bar{P}_0 \in \mathcal{W}(\mathcal{X}, \mathcal{D})$,

$$\mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) = \biguplus_{P_\odot \in \mathcal{L}(\bar{P}_0)} \mathcal{J}(P_\odot).$$

$$\dim \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) = |\mathcal{E}| - \sum_{(x,x') \in \mathcal{D}} |\mathcal{S}_x| + |\mathcal{D}| - |\mathcal{X}|.$$

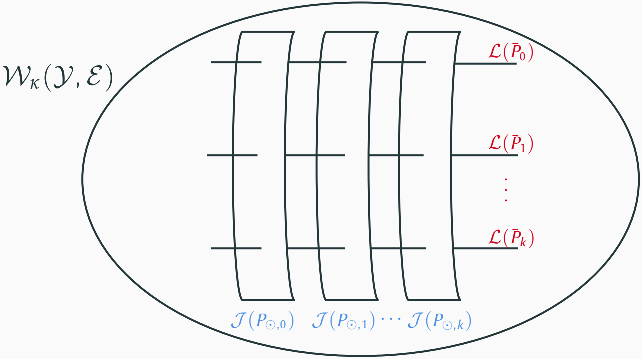
# Foliated manifold of lumpable kernels

**Theorem 14**

*For any fixed* $\bar{P}_0 \in \mathcal{W}(\mathcal{X}, \mathcal{D})$,

$$\mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) = \biguplus_{P_\odot \in \mathcal{L}(\bar{P}_0)} \mathcal{J}(P_\odot).$$

$$\dim \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E}) = |\mathcal{E}| - \sum_{(x,x') \in \mathcal{D}} |\mathcal{S}_x| + |\mathcal{D}| - |\mathcal{X}|.$$



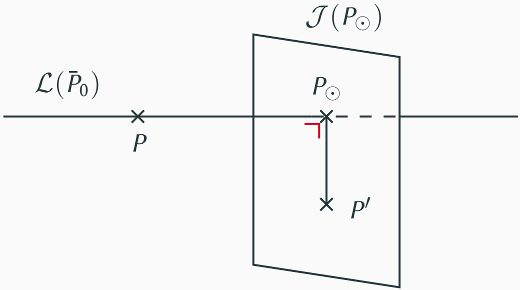
# Application: Leaf projection

**Theorem 15 (Pythagorean identity)**
*Fix* $\bar{P}_0 \in \mathcal{W}(\mathcal{X}, \mathcal{D})$. *Let* $P_\odot, P \in \mathcal{L}(\bar{P}_0)$ *and* $P' \in \mathcal{J}(P_\odot)$.

$$D\left(P \| P'\right) = D\left(P \| P_\odot\right) + D\left(P_\odot \| P'\right),$$

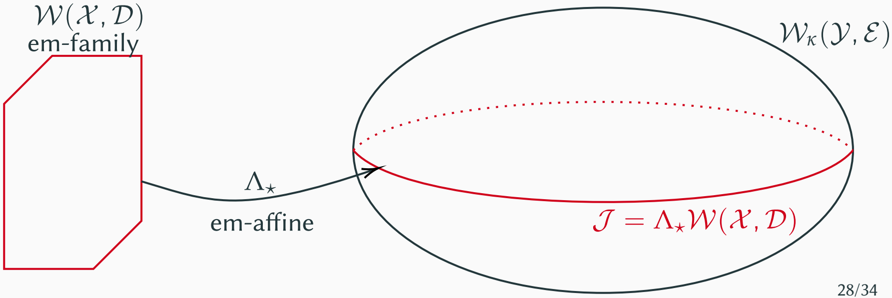
*and* $P_\odot$ *verifies*

$$P_\odot = \underset{P'' \in \mathcal{L}(\bar{P}_0)}{\arg\min}\; D\left(P'' \| P'\right) = \underset{P'' \in \mathcal{J}(P_\odot)}{\arg\min}\; D\left(P \| P''\right).$$



$\mathcal{W}(\mathcal{X}, \mathcal{D})$ em-family, and $\Lambda_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \to \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$, em-affine.

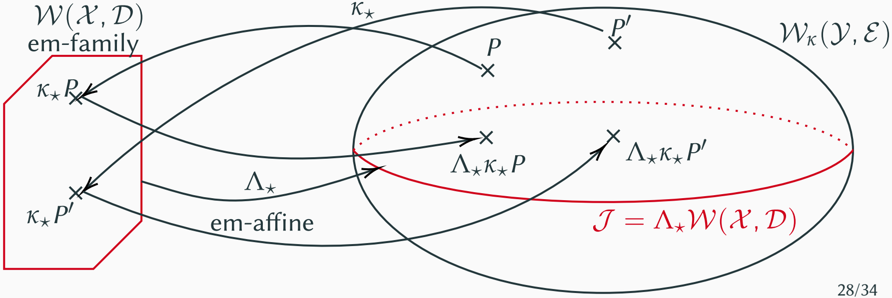
Then $\mathcal{J} \triangleq \{\Lambda_\star \bar{P} \colon \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D})\}$, is em-family,



# Data-processing inequality & m-contraction

$\mathcal{W}(\mathcal{X}, \mathcal{D})$ em-family, and $\Lambda_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \to \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$, em-affine.

Then $\mathcal{J} \triangleq \{\Lambda_\star \bar{P} \colon \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D})\}$, is em-family,

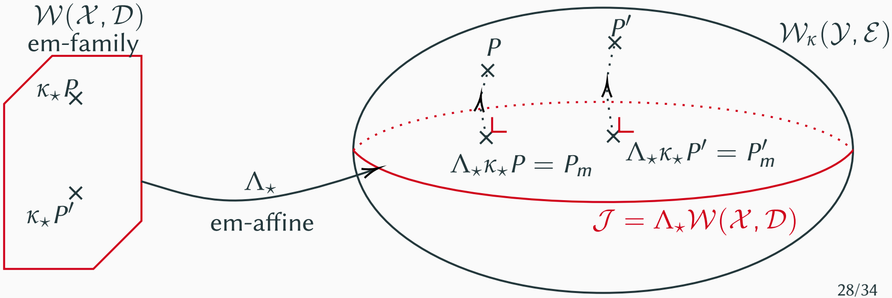


# Data-processing inequality & m-contraction

$\mathcal{W}(\mathcal{X}, \mathcal{D})$ em-family, and $\Lambda_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \to \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$, em-affine.

Then $\mathcal{J} \triangleq \{\Lambda_\star \bar{P} \colon \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D})\}$, is em-family,

For any $P \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$, $P_m \triangleq \arg\min_{\tilde{P} \in \mathcal{J}} D\left(P \| \tilde{P}\right) \overset{\text{Lemma}}{=} \Lambda_\star \kappa_\star P$.



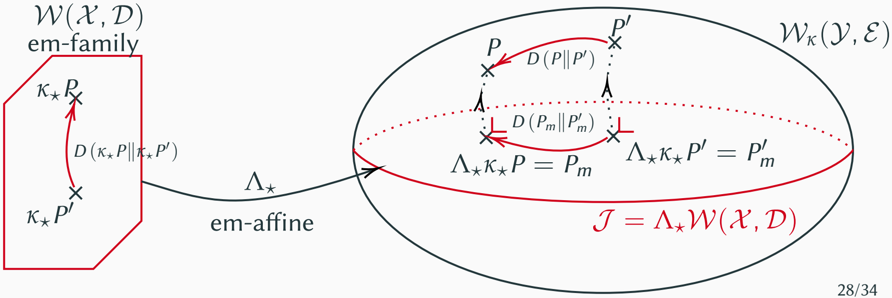
# Data-processing inequality & m-contraction

$\mathcal{W}(\mathcal{X}, \mathcal{D})$ em-family, and $\Lambda_\star \colon \mathcal{W}(\mathcal{X}, \mathcal{D}) \to \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$, em-affine.

Then $\mathcal{J} \triangleq \{\Lambda_\star \bar{P} \colon \bar{P} \in \mathcal{W}(\mathcal{X}, \mathcal{D})\}$, is em-family,

For any $P \in \mathcal{W}_\kappa(\mathcal{Y}, \mathcal{E})$, $P_m \triangleq \arg\min_{\tilde{P} \in \mathcal{J}} D\left(P \| \tilde{P}\right) \overset{\text{Lemma}}{=} \Lambda_\star \kappa_\star P$.

$$D\left(P \| P'\right) \geq D\left(\Lambda_\star \kappa_\star P \| \Lambda_\star \kappa_\star P'\right) = D\left(\kappa_\star P \| \kappa_\star P'\right).$$



**Thank you for listening!**

**e-print: 2203.04575**

# References

Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information Geometry*. Springer, Cham, 2017. doi: https://doi.org/10.1007/978-3-319-56478-4.

L Lorne Campbell. An extended Čencov characterization of the information metric. *Proceedings of the American Mathematical Society*, 98(1):135–141, 1986.

Imre Csiszár, Thomas M. Cover, and Byoung-Seon Choi. Conditional limit theorems under Markov conditioning. *IEEE Transactions on Information Theory*, 33(6):788–801, 1987.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

Jürgen Gärtner. On large deviations from the invariant measure. *Theory of Probability & Its Applications*, 22(1):24–39, 1977.

Masahito Hayashi. Local equivalence problem in hidden Markov model. *Information Geometry*, 2(1):1–42, 2019.

Masahito Hayashi. Information geometry approach to parameter estimation in hidden markov model. *Bernoulli*, 28(1):307–342, 2022.

Masahito Hayashi and Shun Watanabe. Information geometry approach to parameter estimation in Markov chains. *The Annals of Statistics*, 44(4): 1495 – 1535, 2016. doi: 10.1214/15-AOS1420.

Hisashi Ito and Shun'ichi Amari. Geometry of information sources. In *Proceedings of the 11th Symposium on Information Theory and Its Applications (SITA '88)*, pages 57–60, 1988.

John G Kemeny and J Laurie Snell. *Finite Markov chains: with a new appendix" Generalization of a fundamental matrix"*. Springer, New York, 1983.

Guy Lebanon. An extended Čencov-Campbell characterization of conditional information geometry. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, page 341–348, Virginia, 2004. AUAI Press. ISBN 0974903906.

Guy Lebanon. Axiomatic geometry of conditional models. *IEEE Transactions on Information Theory*, 51(4):1283–1294, 2005.

HD Miller. A convexity property in the theory of random variables defined on a finite Markov chain. *The Annals of mathematical statistics*, pages 1260–1270, 1961.

Guido Montúfar, Johannes Rauh, and Nihat Ay. On the Fisher metric of conditional probability polytopes. *Entropy*, 16(6):3207–3233, 2014.

Vrettos Moulos and Venkat Anantharam. Optimal Chernoff and Hoeffding bounds for finite state Markov chains. *arXiv preprint arXiv:1907.04467*, 2019.

Hiroshi Nagaoka. The exponential family of Markov chains and its information geometry. In *The proceedings of the Symposium on Information Theory and Its Applications*, volume 28(2), pages 601–604, 2005.

Kenji Nakagawa and Fumio Kanaya. On the converse theorem in statistical hypothesis testing for Markov chains. *IEEE transactions on information theory*, 39(2):629–633, 1993.

Jun'ichi Takeuchi and Andrew R Barron. Asymptotically minimax regret by Bayes mixtures. In *Proceedings. 1998 IEEE International Symposium on Information Theory (Cat. No. 98CH36252)*, page 318. IEEE, 1998.

Jun'ichi Takeuchi and Tsutomu Kawabata. Exponential curvature of Markov models. In *2007 IEEE International Symposium on Information Theory*, pages 2891–2895. IEEE, 2007.

Jun'ichi Takeuchi and Hiroshi Nagaoka. On asymptotic exponential family of Markov sources and exponential family of Markov kernels, 2017.

Nikolai Nikolaevich Čencov. Algebraic foundation of mathematical statistics. *Series Statistics*, 9(2):267–276, 1978. doi: 10.1080/02331887808801428.

Shun Watanabe and Masahito Hayashi. Finite-length analysis on tail probability for Markov chain and application to simple hypothesis testing. *Ann. Appl. Probab.*, 27(2):811–845, 04 2017. doi: 10.1214/16-AAP1216.

Geoffrey Wolfer and Shun Watanabe. Information geometry of reversible Markov chains. *Information Geometry*, 4(2):393–433, 12 2021. ISSN 2511-2481. doi: 10.1007/s41884-021-00061-7.