Perspectives on Markov Chains: Toward Contraction of Rényi and Csiszár Divergence

Andrew R. Barron

Yale University

Reńyi Centennial

21 June, 2022

Probabilistic Transition from U to V

Probability	Transition Kernel	Induced Distribution
$P = P_U$	$K = P_{V\|U}$	$P^* = P_V$

with mass or density functions

$p(u)$	$p(v\|u)$	$p^*(v) = \sum_u p(v\|u)p(u)$

Joint density: $p(u, v) = p(u)p(v|u)$

Reversed representation: $p(u, v) = p^*(v)p^*(u|v)$

Reversed Transition from V to U

The reversed transition kernel K^* from V to U is expressed by the Bayes reversal rule:

$$p^*(u|v) = p(u, v)/p^*(v)$$

Probability	Transition Kernel	Induced Distribution
$P^* = P_V$	$K^* = P^*_{U\|V}$	$P = P_U$

Gibbs cycle from U to V to $\tilde{U}$

- Application of KK^*, transition kernel K followed by K^*
- Joint probability of U and $\tilde{U}$ in cycle if start with $\pi = P$

$$\begin{aligned} p(u, \tilde{u}) &= \int_v p(u)p(v|u)p^*(\tilde{u}|v) \\ &= \int_v p^*(v)p^*(u|v)p^*(\tilde{u}|v) \\ &= E_V\ p^*(u|V)p^*(\tilde{u}|V) \end{aligned}$$

- Symmetric with equal marginals $\pi = P_U = P_{\tilde{U}}$
- Implies the Gibb's cycle is reversible with invariant distribution $\pi = P$

Other distributions for U

- Distributions Q arise at the start or on a path toward $P = \pi$
- Consider Q with a mass or density ratio with respect to P

$$q(u) = f(u)p(u) \qquad f(u) = \frac{q(u)}{p(u)}$$

- What happens to Q and P and their density ratio when K is applied to each?

$$q^*(v) = \sum_u f(u)p(u)p(v|u)$$

$$p^*(v) = \sum_u p(u)p(v|u)$$

- Ratio obtained for the distributions induced by the forward transition are interestingly found by application of the Bayes reversal

$$f^*(v) = \frac{q^*(v)}{p^*(v)} = \frac{\sum_u f(u)p(u)p(v|u)}{p^*(v)} = \sum_u f(u)p^*(u|v)$$

Relative Entropy between Q and P

- Also called Kullback divergence and information divergence
- This ordinary relative entropy $D(Q||P)$ is $R_1(Q||P)$ in the Rényi relative entropy list on the next page with $\alpha = 1$

$$D(Q||P) = E_Q\left[\log\frac{q(U)}{p(U)}\right] = E_P\left[\frac{q(U)}{p(U)}\log\frac{q(U)}{p(U)}\right]$$

- In literature on functional inequalities, with P fixed, it is sometimes also written, when $E[f] = 1$,

$$\mathrm{Entropy}_P(f) = E[f\log f]$$

or, in unnormalized cases,

$$\mathrm{Entropy}_P(f) = E[f\log f] - E[f]\log E[f],$$

where E denotes expectation with respect to P.

Rényi Relative Entropy between Q and P

Also called the Rényi α divergence.

For $\alpha > 0$ and $\alpha \neq 1$

$$\begin{aligned} R_\alpha(Q||P) &= \frac{1}{\alpha - 1} \log E_P \left(\frac{q(U)}{p(U)} \right)^\alpha \\ &= \frac{1}{\alpha - 1} \log E_Q \left(\frac{q(U)}{p(U)} \right)^{\alpha - 1} \end{aligned}$$

and, for $\alpha = 1$, the $R_1(Q||P)$ is the ordinary relative entropy from the previous page.

Csiszár Divergence between Q and P

- Also called f divergence. For distinction use ϕ in place of f.

- Csiszár Divergence for convex ϕ, here with $\phi(1) = 0$,

$$D_\phi(Q||P) \;=\; E_P\, \phi\left(\frac{q(U)}{p(U)}\right)$$

  General Csiszár Divergence for increasing ψ, convex ϕ, here with $\psi(\phi(1)) = 0$,

$$D_{\psi,\phi}(Q||P) \;=\; \psi\left(E_P\, \phi\left(\frac{q(U)}{p(U)}\right)\right)$$

- The Rényi α divergences arise by building the convex function $\phi(r)$ from r^α.

- The corresponding Csiszár α divergences arise from using $\phi(r)$ equal to $r^\alpha - 1$ and $1 - r^\alpha$, respectively, for $\alpha > 1$ and $\alpha < 1$.

Chi-square Divergence between Q and P

Among Csiszár divergences we find a distinguished role for $\alpha = 2$. In which case one may also take $\phi(r) = \text{square}(r) = (r-1)^2$, leading to the Chi-square divergence

$$D_2(Q||P) = E_P \left(\frac{q}{p} - 1 \right)^2 = E[(f-1)^2]$$

The corresponding Rényi square divergence is

$$R_2 = \log(1 + D_2)$$

Csiszár Divergence Reduction

- Divergence reduction by probabilistic transition
- Also called the Data Processing Inequality
- It expresses what happens to the distance between Q and P when a transition K from U to V is applied to each

$$D_\phi(Q_V || P_V) \leq D_\phi(Q_U || P_U)$$

- Equivalently

$$E_V \phi(f^*(V)) \leq E_U \phi(f(U))$$

Individual Contraction Coefficient

- The individual contraction coefficient is not more than 1
- It is denoted by

$$\rho(K, f) = \rho(\phi, P, K, f)$$

  defined by

$$\rho(K, f) = \frac{D_\phi(Q_V||P_V)}{D_\phi(Q_U||P_U)} = \frac{E_V \phi(f^*(V))}{E_U \phi(f(U))}$$

- It expresses by what fraction the ϕ divergence reduces with transition rule K if we were at density ratio f with respect to P.
- Sometimes ϕ, P, or even K is fixed.

Universal Contraction Coefficient

- The transition K is said to provide strict ϕ divergence contraction if the value is less than 1 for the supremum of the contraction coefficients over all probability density ratios f with respect to P
- The universal contraction coefficient $\rho(K) = \rho(\phi, P, K)$ is, accordingly, defined by

$$\rho(\phi, P, K) = \sup_f \rho(\phi, P, K, f)$$

- The supremum is over all functions $f \geq 0$ with $E_P[f] = 1$.
- Monotonicity: $\rho(\phi, P, KK^*) \leq \rho(\phi, P, K)$.

Markov Chain Convergence

Accordingly, if K is a strict ϕ divergence contraction, then the Gibbs cycle process producing distributions Q_k for U_k at time k, starting from a distribution Q_0 with finite ϕ divergence from P, will converge exponentially fast to P in ϕ divergence with exponent independent of the path. That is, for all integer time steps $k \geq 0$, setting

$$D(k) = D_\phi(Q_k||P)$$

we have

$$D(k) \leq D(0)\,\rho^k$$

where $\rho = \rho(\phi, P, K)$.

Extremality of Square Divergence Contraction

- Let $\text{square}(r) = (r - 1)^2$ providing the Chi-square $D_2(P||Q)$

- Extremal for univ. contraction among Csiszár divergences

- Indeed, for each P, K,

$$\rho(\text{square}, P, K) \leq \rho(\phi, P, K)$$

for all convex $\phi(r)$, twice differentiable at $r = 1$.

Proof of Square Divergence Contraction Extremality

- Proof considers distributions $\tau Q + (1-\tau)P$ on the line between P and Q. These have density ratios $1+\tau(f-1)$ for U and transform to density ratios $1+\tau(f^*-1)$ for V.

- Examine

$$\rho(\phi, P, K, 1+\tau(f-1)) = \frac{E_V \phi(1+\tau(f^*(V)-1))}{E_U \phi(1+\tau(f(U)-1))}$$

- Take the limit as $\tau \to 0$ using two applications of L'Hopital's rule, to get

$$\frac{E_V (f^*(V)-1)^2}{E_U (f(U)-1)^2} = \rho(\text{square}, P, K, f)$$

- observe with the function $g(u) = f(u) - 1$ of mean 0, that the ratio is invariant to rescalings $Cg(u)$ for all $C \neq 0$.

Relating U, V Properties to $U \to V \to \tilde{U}$ Properties

- U to V contraction matches $U, \tilde{U}$ maximal correlation
- One may decompose the chi-square between Q_V and P_V starting from $f = 1 + g$ with $E[g] = 0$. It is $E_V(g^*(V))^2$ expressible as

$$E_V(\sum_u g(u)p^*(u|v))^2$$

Apply the square of sums trick to write this as

$$E_V(\sum_u g(u)p^*(u|v))(\sum_{\tilde{u}} g(\tilde{u})p^*(\tilde{u}|v)) = \sum_{u,\tilde{u}} g(u)g(\tilde{u})p(u,\tilde{u})$$

- This is $COV(g(U), g(\tilde{U}))$. Here we used $p(u,\tilde{u}) = E_V\left[p(u|V)p(\tilde{u})|V\right]$, the stochastic representation of the $u, \tilde{u}$ distribution via V.
- Accordingly, dividing by $E[g^2]$ and taking the maximum, one has

$$\rho(\text{square}, P_U, P_{V|U}) = \sup_{g:Eg=0} \frac{E[g(U)g(\tilde{U})]}{Eg^2}$$

U to V Contraction, Maximal Correlation, or Eigenvalue Characterization

- We have

$$\rho(P_{V|U}) = \sup_{g:Eg=0} \frac{E[g(U)g(\tilde{U})]}{Eg^2} = \sup_{g:Eg=0} \frac{\sum_{u,\tilde{u}} g(u)p(u,\tilde{u})g(\tilde{u})}{||g||_\pi^2}$$

- This endows $\rho(P_{V|U})$ with the additional interpretations familiar in the analysis of $U, \tilde{U}$ joint distributions, including that it is the second largest eigenvalue of $\mathbf{P}$ where $\mathbf{P}$ has entries $p(u,\tilde{u})$, decomposed into functions orthonormal in $L_2(\pi)$. But such eigenvalue representation is not directly illuminating for high-dimensional state space settings.
- Instead we advocate going the other way. If you have a U to $\tilde{U}$ reversible chain, seek an intermediate random variable V that permits the cycle representation $U \to V \to \tilde{U}$.

Transitional Comments on Contraction Strategy

- As we shall see, the U to V contraction is directly amenable to single-letter characterization in multivariate settings, more so than the $U, \tilde{U}$ eigenvalue or maximal correlation.
- Once we single-letterize, then the maximal correlation interpretation becomes helpful on the reduced state space.
- We are motivated by ideas of Yuansi Chen, Ronen Eldan (on foundations of stochastic localization, June 22 arXiv) and Anari, Liu and Oveis Charan (on spectral independence, STOC 19).
- However we use sums of squares of martingale difference identities to get what we think is the most natural single letter characterization for Chi-square rather than the expected product of conditional variance ratio identities used in the above.

Tricks for Multivariate States

- The state space U can be quite large, e.g., $\{-1, 1\}^n$, which has cardinality 2^n.
- Suppose the state has a vector or string representation

$$\underline{u} = (u_1, \ldots, u_n) = u^n.$$

- Then for any function $g(\underline{u})$ of mean $E_P g(\underline{U}) = 0$, there is the sum of uncorrelated martingale difference representation

$$g(\underline{u}) = \sum_{t=1}^{n} g_t(\underline{u})$$

where

$$g_t = g_t(u_t|u^t) = E[g|u^t] - E[g|u^{t-1}].$$

- The $E[g|u^t]$ are the Doob martingale evaluations, averaging g using the conditional distribution of $(U_{t+1}, \ldots, U_n)$ given $U^t = u^t$.

Toward Additive Representation of Chi-square

- Write Chi-square divergences on V (of arbitrary dimension) using the fact that they come from multivariate distributions on $\underline{U} = U^n$. Then $D_2(Q_V || P_V)$ is given by
$$E_V[(\sum_{\underline{u}} g(\underline{u}) p(\underline{u}|V))^2]$$
which is
$$E_V[(\sum_{t=1}^{n} (\sum_{\underline{u}} g_t(\underline{u}) p(\underline{u}|v)))^2]$$
- The interior sums $\sum_{\underline{u}} g_t(\underline{u}) p(\underline{u}|v)$ marginalize to
$$\sum_{u^{t-1}} p^*(u^{t-1}|v) \sum_{u_t} p(u_t|v, u^{t-1}) g_t(u_t, u^{t-1})$$

Toward Additive Representation of Chi-square

- By Cauchy Schwarz, the squares of these interior sums are not more than

$$\sum_{u^{t-1}} p(u^{t-1}|v)\Big(\sum_{u_t} p(u_t|v,u^{t-1})g_t(u_t,u^{t-1})\Big)^2$$

- Averaging it with respect to P_V produces an expectation with respect to the joint distribution of (U^{t-1},V), which permits a reversal of the iterated expectation as

$$E_{U^{t-1}} \int p(v|u^{t-1})\Big(\sum_{u_t} p(u_t|v,u^{t-1})g_t(u_t,u^{t-1})\Big)^2$$

- We recognize the integral as a conditional Chi-square divergence arising from a transition from the single coordinate u_t to V, conditioning on the preceding u^{t-1}.
- Let ρ_1 denote the maximum of such single coordinate contraction coefficients over choices of the conditioning events.

Toward Additive Representation of Chi-square

- Again ρ_1 denotes the maximum of such single coordinate contraction coefficients over choices of the conditioning events.
- So these integrals are not more than ρ_1 times $E_{U_t|U^{t-1}} g_t^2$.
- Taking the expectation and summing over t produces $\rho_1 E g^2$.
- This provides a strategy for demonstrating that

$$\rho(\text{square}, P_{U^n}, P_{V|U^n})$$

is less than or equal to ρ_1.

Study of Contraction from Binary States

- The conditional contraction coefficients arising in the preceding analysis are special cases of the family of binary contraction problems, with

$$u \in \{-1, +1\}$$

- Investigate

$$\rho_1(P_U, P_{V|U}, 1 + g)$$

- In the binary case g with mean 0 must take the form of a multiple of $u - \mu$ where $\mu = E_P[U] = p(1) - p(-1)$.
- All such g have the same individual contraction coefficient, and so these individual contraction coefficients coincide with the universal in the binary case, matching $CORR(U, \tilde{U})$.

The Polarized Case where Contraction Fails

- Again we have

$$\rho_1(P_U, P_{V|U}) = CORR(U, \tilde{U})$$

- Contraction fails when $\tilde{U} = 1$ when $U = 1$

  and $\tilde{U} = -1$ when $U = -1$

Quantifying Favorable Cases for Binary Contraction

- Let $\text{diff}(v) = E[U|v] = p(U=1|v) - p(U=-1|v)$.
- Then

$$\rho_1(P_U, P_{V|U}) = \frac{E_V(\sum_u p(u|v)(u-\mu))^2}{VAR(U)}$$

- It is seen to equal

$$\frac{E[\text{diff}^2(V) - \mu^2]}{1-\mu^2}$$

- which is not more than

$$E[\text{diff}^2(V)]$$

Quantifying Favorable Cases for Binary Contraction

- We have
$$\rho_1(P_U, P_{V|U}) \leq E[\mathrm{diff}^2(V)].$$
- So we look to the distribution of $\mathrm{diff}(V)$ in $[-1, 1]$ induced by the distribution on V (condition distributions in the application to the vector case).
- Problematic if it is bimodally spiked at -1 and $+1$. Any mass away from the extremes is sufficient to produce contraction.
- For instance to have $\rho_1 \leq 1 - \delta$ it is enough that this distribution assigns probability at least α to the cases with $\mathrm{diff}(V)$ at least δ/α away from ± 1.