Distance-Divergence Inequalities

Katalin Marton

Alfréd Rényi Institute of Mathematics
of the Hungarian Academy of Sciences
Motivation

- To find a simple proof of the **Blowing-up Lemma**, proved by Ahlswede, Gács and Körner, with the aim to prove **Strong Convereses** in Shannon theory;

- Ornstein’s **Copying Lemma** in the proof of his **Isomorphism Theorem**;

- My experience with **Single-Letter Characterization techniques** in Shannon theory.
What does measure concentration mean?

Enlargements and distances of sets

Notation

\((X, d, q)\): a metric probability space. More precisely,

Assume:

- \(X\) is a Polish space, with the Borel \(\sigma\)-algebra;
- \(d\) is a Borel measurable distance on \(X\);
- \(q\) is a Borel probability measure on \(X\);
- Subsets of \(X\) are assumed to be measurable.
**Notation**

For $A \subset \mathcal{X}$ and $\kappa > 0$:

The $\kappa$-enlargement (or $\kappa$-neighborhood) of the set $A$ is

$$[A]_\kappa = \left\{ x \in \mathcal{X} : \exists \ y \in A, \ d(x, y) \leq \kappa \right\}.$$

**Informal definition of Measure Concentration:**

The a metric probability space $(\mathcal{X}, d, q)$ satisfies the measure concentration property if:

For every $\varepsilon > 0$ there exists a $\kappa = \kappa(\varepsilon)$ such that

$$q(A) \geq \varepsilon \implies q([A]_\kappa) \geq 1 - \varepsilon, \ \forall A \subset \mathcal{X},$$

and $\kappa(\varepsilon)$ can be overbounded in a non-trivial way.
\[ d(A, [A]_\kappa^c) = \kappa \]

\[ B = [A]_\kappa^c \]
By the equality

\[ d(A, [A]^c_\kappa) = \kappa : \]

---

**Equivalent definition**

*The a metric probability space \((\mathcal{X}, d, q)\) satisfies the measure concentration property if:*

*There is a function \(\kappa : \mathbb{R}_+ \rightarrow \mathbb{R}_+\) such that:*

*For every \(A, B \subset \mathcal{X}\)*

\[ d(A, B) \leq \kappa(q(A)) + \kappa(q(B)), \quad \forall A, B \subset\mathcal{X}, \]

*and \(\kappa\) can be overbounded in a non-trivial way.*
Sub-Gaussian measure concentration

\[ \kappa(\varepsilon) = c(q) \cdot \sqrt{\log \frac{1}{\varepsilon}} \quad \iff \quad \varepsilon(\kappa) = \exp\left(-\left(\frac{\kappa}{c(q)}\right)^2\right) \]

I.e.:

\[ d(A, B) \leq c(q) \cdot \left[ \sqrt{\log \frac{1}{q(A)}} + \sqrt{\log \frac{1}{q(B)}} \right], \quad \forall A, B \subset X. \]
Example: Hamming distance, C. McDiarmid 1989, Marton 1996

- $\mathcal{X}^n$: $n$-th power of a Borel space $\mathcal{X}$;
- $\delta_n(x^n, y^n) = \sum_{i=1}^{n} \delta(x_i, y_i)$: Hamming distance on $\mathcal{X}^n$;
- $x^n = (x_1, x_2, \ldots, x_n)$;
- $q^n = \prod_{i=1}^{n} q_i$: product measure on $\mathcal{X}^n$.

Then, taking neighborhood with respect to Hamming distance:

$$q^n([A]_\kappa) \geq 1 - e^{-2n\left(\frac{\kappa}{n} - \sqrt{\frac{1}{2n} \cdot \log \frac{1}{q^n(A)}}\right)^2}, \quad \forall \kappa, A \subset \mathcal{X}^n.$$
Cont’d

\[ q^n([A]_\kappa) \geq 1 - \\
- \exp \left( -2n \left( \frac{\kappa}{n} - \sqrt{\frac{1}{2n} \cdot \log \frac{1}{q^n(A)}} \right)^2 \right), \quad \forall \kappa, A \subset \mathcal{X}^n. \]

Equivalently:

\[ \delta_n(A, B) \leq \sqrt{\frac{1}{2n} \cdot \log \frac{1}{q^n(A)}} + \sqrt{\frac{1}{2n} \cdot \log \frac{1}{q^n(B)}}, \quad \forall A, B \subset \mathcal{X}^n \]

Thus, for Hamming distance \( \delta_n \)

\[ \kappa(\varepsilon) = \sqrt{\frac{1}{2n} \cdot \log \frac{1}{\varepsilon}}. \]
Cont’d

\[ q^n([A]_\kappa) \geq 1 - \exp\left(-2n\left(\frac{\kappa}{n} - \sqrt{\frac{1}{2n} \cdot \log \frac{1}{q^n(A)}}\right)^2\right), \quad \forall \kappa, A \subset \mathcal{X}^n. \]

**Equivalently:**

\[ \delta_n(A, B) \leq \sqrt{\frac{1}{2n} \cdot \log \frac{1}{q^n(A)}} + \sqrt{\frac{1}{2n} \cdot \log \frac{1}{q^n(B)}}, \quad \forall A, B \subset \mathcal{X}^n \]

**Thus, for Hamming distance** \( \delta_n \)

\[ \kappa(\varepsilon) = \sqrt{\frac{1}{2n} \cdot \log \frac{1}{\varepsilon}}. \]
Example: Euclidean distance, Talagrand 1996

- $\mathbb{R}^n$: $n$-dimensional Euclidean space;
- $e_n$: Euclidean distance;
- $q^n = \prod_{i=1}^{n} q_i = \prod_{i=1}^{n} \exp(-V_i)$: product density on $\mathbb{R}^n$.

Assume:
$V_i : \mathbb{R} \to \mathbb{R}$ is strictly uniformly convex:

$$V_i''(x) \geq \rho > 0, \quad \forall x.$$  

Then:

$$e_n(A, B) \leq \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{q^n(A)}} + \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{q^n(B)}}, \quad \forall A, B \subset \mathcal{X}^n.$$
Example: Euclidean distance, Talagrand 1996

- \( \mathbb{R}^n \): \textit{n-dimensional Euclidean space};

- \( e_n \): \textit{Euclidean distance};

- \( q^n = \prod_{i=1}^{n} q_i = \prod_{i=1}^{n} \exp(-V_i) \): \textit{product density on} \( \mathbb{R}^n \).

Assume:
\( V_i : \mathbb{R} \rightarrow \mathbb{R} \) \textit{is strictly uniformly convex}:

\[
V_i''(x) \geq \rho > 0, \quad \forall x.
\]

Then:

\[
e_n(A, B) \leq \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{q^n(A)}} + \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{q^n(B)}}, \quad \forall A, B \subset X^n.
\]
Cont’d

Assume:

\( V_i : \mathbb{R} \to \mathbb{R} \) is strictly uniformly convex:

\[ V_i''(x) \geq \rho > 0, \quad \forall x. \]

Then:

\[ e_n(A, B) \leq \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{q^n(A)}} + \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{q^n(B)}}, \quad \forall A, B \subset \mathcal{X}^n. \]

Thus, for Euclidean distance, and for product density

\[ q^n = \prod_{i=1}^{n} \exp(-V_i) \text{ satisfying } V_i'' \geq \rho > 0, \]

\[ \kappa(\varepsilon) = \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{\varepsilon}}, \quad \text{independently of } n. \]
Assume:

\[ V_i : \mathbb{R} \to \mathbb{R} \text{ is strictly uniformly convex:} \]

\[ V_i''(x) \geq \rho > 0, \quad \forall x. \]

Then:

\[ e_n(A, B) \leq \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{q^n(A)}} + \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{q^n(B)}}, \quad \forall A, B \subset \mathcal{X}^n. \]

Thus, for Euclidean distance, and for product density

\[ q^n = \prod_{i=1}^{n} \exp(-V_i) \text{ satisfying } V_i'' \geq \rho > 0, \]

\[ \kappa(\varepsilon) = \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{\varepsilon}}, \quad \text{independently of } n. \]
Distance-Divergence Inequalities: A way to prove measure concentration

Notation

\[ \mathbb{P}(\mathcal{X}) : \text{the set of probability measures on } \mathcal{X}. \]

\[ q \text{ is a distinguished element of } \mathbb{P}(\mathcal{X}). \]

Definition: set of Couplings \( \Pi(p, r) \)

For \( p, r \in \mathbb{P}(\mathcal{X}) \):

\[ \Pi(p, r) \triangleq \]

set of all couplings of \( p \) and \( r \)

\[ \triangleq \]

set of all joint probability distributions \( \pi = \mathcal{L}(Y, Z) \) on \( \mathcal{X} \times \mathcal{X} \), with marginals:

\[ p = \mathcal{L}(Y) \quad \text{and} \quad r = \mathcal{L}(Z). \]
**Definition: Wasserstein distance**

Fix a $\beta$, $1 \leq \beta \leq 2$. For $p, r \in \mathcal{P}(\mathcal{X})$:

$$W_\beta(p, r) = W_\beta^{(d)}(p, r)$$

$$\triangleq \inf_{\pi \in \Pi(p, r)} \left\{ \left( \mathbb{E}_\pi d(Y, Z)^\beta \right)^{1/\beta} : \mathcal{L}(Y) = p, \quad \mathcal{L}(Z) = r \right\},$$

$W_\beta(p, r)$ is the Wasserstein distance, of order $\beta$, of $p$ and $r$.

**Claim**

$W_\beta$ is a distance on $\mathcal{P}(\mathcal{X})$. (May take the value $\infty$.)
Definition: Wasserstein distance

Fix a $\beta$, $1 \leq \beta \leq 2$. For $p, r \in \mathbb{P}(\mathcal{X})$ :

$$W_\beta(p, r) = W_\beta^{(d)}(p, r)$$

$$\triangleq \inf_{\pi \in \Pi(p, r)} \left\{ \left( \mathbb{E}_\pi d(Y, Z)^\beta \right)^{1/\beta} : \mathcal{L}(Y) = p, \quad \mathcal{L}(Z) = q \right\},$$

$W_\beta(p, r)$ is the Wasserstein distance, of order $\beta$, of $p$ and $r$.

Example: Variational distance

$\delta$: Kronecker’s distance: $\delta(x, y) = 1$ for $x \neq y, \delta(x, x) = 0$.

$$W_1^{(\delta)}(p, r) \triangleq |p - r|_{TV} = \min \mathbb{E}_\pi \delta(Y, Z)$$

$$= \min \Pr_\pi \{Y \neq Z\} = \max_{A \subset \mathcal{X}} (p(A) - r(A)),$$

where $\pi = \mathcal{L}(Y, Z) \in \Pi(p, r)$. 
Definition: Wasserstein distance

Fix a $\beta$, $1 \leq \beta \leq 2$. For $p, r \in \mathbb{P}(X)$:

$$W_\beta(p, r) = W^{(d)}_\beta(p, r) \triangleq \inf_{\pi \in \Pi(p, r)} \left\{ \left( \mathbb{E}_\pi d(Y, Z)^\beta \right)^{1/\beta} : \mathcal{L}(Y) = p, \quad \mathcal{L}(Z) = q \right\}.$$

$W_\beta(p, r)$ is the Wasserstein distance, of order $\beta$, of $p$ and $r$.

Example: Variational distance

$\delta$: Kronecker’s distance: $\delta(x, y) = 1$ for $x \neq y$, $\delta(x, x) = 0$.

$$W_1^{(\delta)}(p, r) \triangleq |p - r|_{TV} = \min \mathbb{E}_\pi \delta(Y, Z) = \min Pr\pi\{Y \neq Z\} = \max_{A \subseteq \mathcal{X}} (p(A) - r(A)),$$

where $\pi = \mathcal{L}(Y, Z) \in \Pi(p, r)$.
Definition: (Informational) Divergence

For $p, q \in \mathbb{P}(\mathcal{X})$:

$$D(p||q) \triangleq \int_{\mathcal{X}} \log \left( \frac{dp}{dq} \right) dp \quad \text{if } p \ll q; \quad +\infty \quad \text{otherwise.}$$

Definition: Distance-Divergence Inequality

$q \in \mathbb{P}(\mathcal{X})$ satisfies a distance-divergence inequality of order $\beta$, with constant $\rho$, if:

$$W_{\beta}(p, q) \leq \sqrt{\frac{2}{\rho} \cdot D(p||q)} \quad \text{for all } p \in \mathbb{P}(\mathcal{X}).$$

Most important cases: $\beta = 1$ and $\beta = 2$. 
Definition: Distance-Divergence Inequality

$q \in \mathbb{P}(\mathcal{X})$ satisfies a distance-divergence inequality of order $\beta$, with constant $\rho$, if:

$$W_\beta(p, q) \leq \sqrt{\frac{2}{\rho} \cdot D(p||q)} \quad \text{for all } p \in \mathbb{P}(\mathcal{X}).$$

Example: Pinsker-Csiszár-Kullback inequality

$$|p - q|_{TV} \leq \sqrt{\frac{1}{2} \cdot D(p||q)} \quad \forall p, q \in \mathbb{P}(\mathcal{X}).$$

This distance-divergence inequality holds with the same constant $\rho = 4$ for every $q$. 
Definition: Distance-Divergence Inequality

\( q \in \mathbb{P}(\mathcal{X}) \) satisfies a distance-divergence inequality of order \( \beta \), with constant \( \rho \), if:

\[
W_\beta(p, q) \leq \sqrt{\frac{2}{\rho} \cdot D(p\|q)} \quad \text{for all} \quad p \in \mathbb{P}(\mathcal{X}).
\]

Example: Pinsker-Csiszár-Kullback inequality

\[
|p - q|_{TV} \leq \sqrt{\frac{1}{2} \cdot D(p\|q)} \quad \forall p, q \in \mathbb{P}(\mathcal{X}).
\]

This distance-divergence inequality holds with the same constant \( \rho = 4 \) for every \( q \).
Notation

\( W_2 \):

Wasserstein distance on \( \mathbb{P}(\mathbb{R}) \) from the Euclidean distance, i.e.:

For densities \( p \) and \( r \) on \( \mathbb{R} \)

\[
W_2(p, r) = W_2^{(e)}(p, r) = \inf_{\pi \in \Pi(p, r)} \left\{ \left( \mathbb{E}_{\pi} |Y - Z|^2 \right)^{1/2} : \mathcal{L}(Y) = p, \; \mathcal{L}(Z) = q \right\}.
\]
M. Talagrand’s Theorem on $\mathbb{R}$, 1996

Assume:

$q(x) = \exp(-V(x))$ strictly uniformly log-concave density on $\mathbb{R}$, i.e.,

$$V''(x) \geq \rho > 0, \quad \forall x.$$  

E.g., for $q$ normal, $\rho = 1/\text{Variance}$. Then:

$$W_2(p, r) \leq \sqrt{\frac{2}{\rho}} \cdot D(p\|q), \quad \forall p \in \mathbb{P}(\mathbb{R}).$$
The proof of Talagrand’s Theorem is based on the unique monotone map $T$ that takes the measure $p$ on $\mathbb{R}$ to the measure $q$ (provided both measures are absolutely continuous):

$$\int_{-\infty}^{x} q(t)dt = \int_{-\infty}^{T(x)} p(t)dt.$$ 

A monotone map taking a given density $p$ to the density $q$ exists also on $\mathbb{R}^n$. (Proved, e.g., by Y. Brenier 1991 or R. McCann 1994.)
Talagrand’s Theorem on $\mathbb{R}^n$

Assume:

$q(x) = \exp(-V(x))$: strictly uniformly log-concave density on $\mathbb{R}^n$:

$$\text{Hess}V(x) \geq \rho \cdot \text{Id} \quad \forall x, \quad \rho > 0.$$ 

Then:

$$W_2(p, q) \leq \sqrt{\frac{2}{\rho} \cdot D(p||q)} \quad \forall p \in \mathcal{P}^0(\mathbb{R}).$$

$W_2$ denotes the Wasserstein distance on $\mathcal{P}(\mathbb{R}^n)$ derived from the Euclidean distance.

Proved e.g. by S. Bobkov - M. Ledoux 2000, F. Otto - C. Villani 2000, D. Cordero-Erausquin 2002...
Lemma: Distance-Divergence Inequality implies Measure Concentration, Marton 1986, 1996

\((\mathcal{X}, d, q)\): metric probability space,

\(W^{(d)}_\beta\) Wasserstein distance of order \(\beta\) associated with \(d\).

Then:

\[
W^{(d)}_\beta(p, q) \leq \sqrt{\frac{2}{\rho}} \cdot D(p||q), \quad \forall p \in \mathbb{P}(\mathcal{X})
\]

\[
d(A, B) \leq \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{q(A)}} + \sqrt{\frac{2}{\rho} \cdot \log \frac{1}{q(B)}}, \quad \forall A, B \subset \mathcal{X}.
\]
Proof

Define

\[ p = q_{|A} : \quad p(C) = \frac{q(A \cap C)}{q(A)}, \]
\[ r = q_{|B} : \quad r(C) = \frac{q(B \cap C)}{q(B)}. \]

Triangle inequality \quad + \quad Distance-Divergence inequality:

\[ d(A, B) \leq W_\beta(p, r) \leq W_\beta(p, q) + W_\beta(q, r) \]
\[ \leq \sqrt{\frac{2}{\rho} \cdot D(p||q)} + \sqrt{\frac{2}{\rho} \cdot D(r||q)} \quad \forall \beta \in [1, 2]. \]
d(A, B) ≤ \sqrt{\frac{2}{\rho} \cdot D(p||q)} + \sqrt{\frac{2}{\rho} \cdot D(r||q)}.

Further:
\[
\frac{p(x)}{q(x)} = \frac{1}{q(A)} \quad p - almost everywhere
\]

⇒

\[
D(p||q) = \log \frac{1}{q(A)},
\]

and similarly

\[
D(r||q) = \log \frac{1}{q(B)}.
\]
Product spaces

Definition: Distances in Product Spaces

Given the metric \( d \) on \( X \), define distances on \( X^n \):

\[
d_{\beta,n}(x^n, y^n) = \left( \sum_{i=1}^{n} d(x_i, y_i)^\beta \right)^{1/\beta}, \quad 1 \leq \beta \leq 2.
\]

**Hamming distance:** from Kronecker’s distance, with \( \beta = 1 \),

**Euclidean distance:** from one-dimensional Euclidean distance, with \( \beta = 2 \).
The distance $d_{\beta,n}$ on $\mathcal{X}^n$ gives rise to the Wasserstein distance $W_{\beta,n} = W^{(d)}_{\beta,n}$ on $\mathcal{P}(\mathcal{X}^n)$:

$$W_{\beta,n}(p^n, r^n) = W^{(d)}_{\beta,n}(p^n, r^n)$$

$$= \inf_{\pi \in \Pi(p^n, r^n)} \left\{ \left( \mathbb{E}_\pi \sum_{i=1}^n d(Y_i, Z_i)^\beta \right)^{1/\beta} : \right\}$$

$L(Y^n) = p^n, \ L(Z^n) = r^n$.

\[ q^n = \prod_{i=1}^{n} q_i : \text{product measure on } \mathcal{X}^n. \text{ Then:} \]

\[ W_\beta(p, q_i) \leq \sqrt{\frac{2}{\rho}} \cdot D(p||q_i), \quad \forall i, p \in \mathbb{P}(\mathcal{X}) \]

\[ \implies \]

\[ W_{\beta,n}(p^n, q^n) \leq \sqrt{\frac{2n^{2/\beta-1}}{\rho}} \cdot D(p^n||q^n), \quad \forall p^n \in \mathbb{P}(\mathcal{X}^n). \]

Thus:

For Hamming distance: \( n \)-fold product measures satisfy a distance-divergence inequality with the same constant \( 4/n \).

For Euclidean distance: \( n \)-fold product measures satisfy distance-divergence inequalities with the smallest of the constants valid for the factors, independently of \( n \).

\[ q^n = \prod_{i=1}^{n} q_i : \text{product measure on } \mathcal{X}^n. \text{ Then:} \]

\[ W_\beta(p, q_i) \leq \sqrt{\frac{2}{\rho} \cdot D(p\| q_i)}, \quad \forall i, p \in \mathbb{P}(\mathcal{X}) \]

\[ \Rightarrow \]

\[ W_{\beta, n}(p^n, q^n) \leq \sqrt{\frac{2n^{2/\beta - 1}}{\rho} \cdot D(p^n\| q^n)}, \quad \forall p^n \in \mathbb{P}(\mathcal{X}^n). \]

Thus:

For Hamming distance: \( n \)-fold product measures satisfy a distance-divergence inequality with the same constant \( 4/n \).

For Euclidean distance: \( n \)-fold product measures satisfy distance-divergence inequalities with the smallest of the constants valid for the factors, independently of \( n \).
Proof

From couplings between $q_i$ and the conditional distributions $p_i(\cdot | y^{i-1}) \triangleq \mathcal{L}(Y_i | Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1})$

can construct

good couplings between measures $p^n$ and $q^n$ on $\mathcal{X}^n$:

Successively define

$$\mathcal{L}(Y_i, X_i | Y_i^{i-1} = y_i^{i-1}, X_i^{i-1} = x_i^{i-1})$$

$\triangleq$ best coupling of $p_i(\cdot | y_i^{i-1})$ and $q_i$. 
We have defined a joint distribution \( \pi = \mathcal{L}(Y^n, X^n) \) with marginals \( p^n = \mathcal{L}(Y^n), q^n = \mathcal{L}(X^n) \).

\[
\mathbb{E}_\pi \delta_n(Y^n, X^n) = \sum_{i=1}^{n} \Pr_{\pi} \{ Y_i \neq X_i \} \\
= \sum_{i=1}^{n} \mathbb{E} \{ Y_i \neq X_i | Y^{i-1} \} \leq \sum_{i=1}^{n} \sqrt{\frac{1}{2} \cdot D(Y_i | Y^{i-1} \|| q_i(\cdot | Y^{i-1}))} \\
= \sum_{i=1}^{n} \sqrt{\frac{1}{2} \cdot D(Y_i | Y^{i-1} \|| q_i)} \\
\leq \sqrt{\frac{n}{2} \cdot \sum_{i=1}^{n} D(Y_i | Y^{i-1} \|| q_i)} \\
= \sqrt{\frac{n}{2} \cdot D(p^n || q^n)}. 
\]
S. Bobkov and Götze’s theorem, 1999

\((\mathcal{X}, d, q): \text{metric probability space.}\)

The following two statements are equivalent:

\[(i) \quad W_1^{(d)}(p, q) \leq \sqrt{\frac{2}{\rho} \cdot D(p||q)} \quad \text{for all} \quad p \in \mathbb{P}(\mathcal{X}),\]

\[(ii) \quad \int_{\mathcal{X}} e^{tf(x)} dq(x) \leq \exp\left(\frac{t^2}{2\rho} + t \cdot \mathbb{E}_q f\right)\]

for all Lipschitz functions \(f: \mathcal{X} \to \mathbb{R}\) with Lipschitz coefficient 1, and all \(t > 0\).
Remark

Bobkov and Götze actually proved that:

For any fixed function \( f : \mathcal{X} \to \mathbb{R} \):

\[
\int_{\mathcal{X}} e^{tf(x)} dq(x) \leq \exp\left( \frac{t^2}{2\rho} + t \cdot \mathbb{E}_q f \right)
\]

\[
\Rightarrow
\]

\[
|\mathbb{E}_p f - \mathbb{E}_q f| \leq \sqrt{\frac{2}{\rho}} \cdot D(p \| q) \quad \text{for all} \quad p \in \mathbb{P}(\mathcal{X}).
\]
(\(\mathcal{X}, d, q\)) : \text{metric probability space}, \quad q^n : \text{n-th power of } q.

The following two statements are equivalent:

- \((i)\) \(W_2^{(d)}(p, q) \leq \sqrt{\frac{2}{\rho}} \cdot D(p||q)\) for all \(p \in \mathbb{P}(\mathcal{X})\),

- \((ii)\) \(q^n \left\{ x^n : f(x^n) \geq \mathbb{E}f + r \right\} \leq \exp\left(\frac{-\rho r^2}{2}\right)\)

for all \(n\), all Lipschitz functions \(f : \mathcal{X}^n \mapsto \mathbb{R}\) with Lipschitz coefficient 1 (with respect to distance \(d_{2,n}\)), and all \(r > 0\).
Measure Concentration for Non-product Measures: Contracting Markov Chains, Marton 1996

Assume:

$q_1$: probability measure on the Borel space $\mathcal{X}$;

$q_i(\cdot|x)$: Markov kernels, $\mathcal{X} \to \mathcal{X}$,

Contracting:

$$|q_i(\cdot|x) - q_i(\cdot|y)|_{TV} \leq (1 - \beta), \quad \beta > 0, \quad \forall x, y;$$

$q^n$ on $\mathcal{X}^n$:

$$q^n(x^n) \triangleq q_1(x_1) \cdot \prod_{i=2}^{n} q_i(x_i|x_{i-1}), \quad x^n \in \mathcal{X}^n.$$

Then:

$$W_{1,n}^{(\delta)}(p^n, q^n) \leq \frac{1}{\beta} \cdot \sqrt{\frac{1}{2n} \cdot D(p^n||q^n)}, \quad \forall p^n \in \mathcal{P}(\mathcal{X}).$$
Measure Concentration for Non-product Measures

\( \mathcal{X}^n \): \( n \)-th power of a Borel space \( \mathcal{X} \);
\( q^n \): Borel probability measure on \( \mathcal{X}^n \);
\( X^n \): random sequence with \( \mathcal{L}(X^n) = q^n \);
\( p^n \): another probability measure on \( \mathcal{X}^n \);
\( Y^n \): random sequence in \( \mathcal{X}^n \), \( \mathcal{L}(Y^n) = p^n \).

Notation

For \( y^n \in \mathcal{X}^n \) and \( 1 \leq i \leq n \):

- \( y^i \triangleq (y_1, y_2, \ldots, y_i) \) and \( y_i^n \triangleq (y_{i+1}, y_{i+2}, \ldots, y_n) \);
- \( p_i(\cdot | y^{i-1}) \triangleq \mathcal{L}(Y_i | Y^{i-1} = y^{i-1}) \);
- \( p_i^n(\cdot | y^i) \triangleq \mathcal{L}(Y_i^n | Y^i = y^i) \);
Definition: Measures admitting coupling with bounded distance

$q^n$ admits coupling with distance bounded by constant $C$ if:

For every $i \leq n - 1$ and every pair of sequences $y^i, z^i \in \mathcal{X}^i$ such that

$$y^{i-1} = z^{i-1}, \quad y_i \neq z_i,$$

there exists a coupling of $q^n_i (\cdot | y^i)$ and $q^n_i (\cdot | z^i)$:

$$\pi^n_i (\cdot | y^i, z^i) = \mathcal{L}(Y^n_i, Z^n_i | Y^i = y^i, Z^i = z^i)$$

satisfying

$$\mathbb{E}_{\pi^n_i} \left\{ \sum_{j=i+1}^{n} \delta(Y_j, Z_j) \mid y^i, z^i \right\} \leq C.$$
Theorem: Measure Concentration for random processes, Marton 1998

Assume:

$q^n$ admits coupling with distance bounded by the constant $C$.

Then

$$W_{1,n}^{(\delta)}(p^n, q^n) \leq (C + 1) \cdot \sqrt{\frac{n}{2} \cdot D(p^n || q^n)}.$$

Thus:

For measures admitting couplings with bounded distance, the distance-divergence inequality is worse by a constant factor (only) than the inequality for product measures.

The Theorem applies for segments of sufficiently mixing stationary random processes.
Measure Concentration for Gibbs Measures

Notation

- $q^n$: Borel probability measure on $\mathcal{X}^n$;
- $X^n$: random sequence, $\mathcal{L}(X^n) = q^n$;
- For $y^n \in \mathcal{X}^n$: $\bar{y}_i \triangleq (y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$;
- $Q_i(\cdot | \bar{y}_i) \triangleq \mathcal{L}(X_i|\bar{X}_i = \bar{y}_i)$.

The conditional distributions $Q_i(\cdot | \bar{y}_i)$ are called the local specifications of $q^n$. 
(Non-standard) Definition: Gibbs Measures

$q^n$, when defined by the local specifications $Q_i(\cdot | \bar{y}_i)$, is called a Gibbs measure.
Definition: Dobrushin’s Uniqueness Condition

$q^n$ satisfies Dobrushin’s uniqueness condition if:
there exist numbers $a_{i,k} \geq 0$, $(i, k \in [1, n], i \neq k)$, such that:

(i) For any $i, k, i \neq k$, and any $\bar{y}_k, \bar{z}_k \in \mathcal{X}^{n-1}$, differing only in the $i$-th coordinate:

$$|Q_k(\cdot|\bar{y}_k) - Q_k(\cdot|\bar{z}_k)|_{TV} \leq a_{i,k};$$

(ii) For the matrix $A \triangleq (a_{i,k})_{i,k=1}^n$, where $a_{i,i} = 0$, we have

$$\|A\|_\infty = \max_k \sum_{i=1}^n a_{i,k} < 1.$$

The numbers $a_{i,k}$ are called Dobrushin’s interdependence coefficients.
Definition: Gibbs Sampler

**Gibbs sampler** associated with the local specifications $Q_i(\cdot | \bar{y}_i)$: Markov chain with state space $\mathcal{X}^n$ and transition law $\Gamma(z^n | y^n)$ defined as follows:

Given $y^n \in \mathcal{X}^n$,

(i) select a random index $i \in [1, n]$ independently of $y^n$ and uniformly distributed;

(ii) Set

$$\Gamma(z^n | y^n) = \delta_{\bar{y}_i, \bar{z}_i} \cdot Q_i(z_i | \bar{y}_i).$$

Dobrushin’s uniqueness condition implies that the Gibbs sampler $\Gamma$ is a contraction with respect to the distance $W_{1,n}^{(\delta)}$:

$$W_{1,n}^{(\delta)}(p^n \Gamma, r^n \Gamma) \leq \left( 1 - \frac{1 - \|A\|_{\infty}}{n} \right) \cdot W_{1,n}^{(\delta)}(p^n, r^n).$$
**Definition: Gibbs Sampler**

**Gibbs sampler associated with the local specifications** $Q_i(\cdot|\bar{y}_i)$:  
**Markov chain with state space** $\mathcal{X}^n$ **and transition law** $\Gamma(z^n|y^n)$ **defined as follows:**

**Given** $y^n \in \mathcal{X}^n$,

1. **(i) select a random index** $i \in [1, n]$ **independently of** $y^n$ **and uniformly distributed;**
2. **(ii) Set**

$$\Gamma(z^n|y^n) = \delta_{\bar{y}_i, \bar{z}_i} \cdot Q_i(z_i|\bar{y}_i).$$

Dobrushin’s uniqueness condition implies that the Gibbs sampler $\Gamma$ **is a contraction with respect to the distance** $W_{1,n}^{(\delta)}$:

$$W_{1,n}^{(\delta)}(p^n \Gamma, r^n \Gamma) \leq \left(1 - \frac{1 - \|A\|_{\infty}}{n}\right) \cdot W_{1,n}^{(\delta)}(p^n, r^n).$$
Theorem, Li Ming Wu 2006

If $q^n$ satisfies Dobrushin’s uniqueness condition then

$$W^{(\delta)}_{1,n}(p^n, q^n) \leq \frac{1}{1 - \|A\|_{\infty}} \sqrt{\frac{n}{2}} \cdot D(p^n || q^n)$$

for all $p^n \in \mathbb{P}(\mathcal{X})^n$.

Remark

A similar theorem holds for Gibbs samplers with state space $\mathbb{R}^n$ and $W^{(e)}_{2,n}$. (With the appropriate modification of Dobrushin’s uniqueness condition.)

Theorem, Li Ming Wu 2006

If $q^n$ satisfies Dobrushin’s uniqueness condition then

$$W_{1,n}^{(\delta)}(p^n, q^n) \leq \frac{1}{1 - \|A\|_\infty} \sqrt{\frac{n}{2} \cdot D(p^n || q^n)}$$

for all $p^n \in \mathbb{P}(X)^n$.

Remark

A similar theorem holds for Gibbs samplers with state space $\mathbb{R}^n$ and $W_{2,n}^{(e)}$. (With the appropriate modification of Dobrushin’s uniqueness condition.)

Theorem: The Gibbs sampler Decreases Divergence by a rate < 1, Marton, in preparation

Assume

(i) $\alpha \triangleq \inf_{x^n,i} Q_i(x_i|x_{\bar{i}}) > 0$;
(ii) $q^n$ satisfies the $L_2$-version of Dobrushin's uniqueness condition:

For the matrix of Dobrushin's interdependence coefficients

$$\|A\|_2 < 1.$$

Then:

$$D(p^n \| q^n) \leq \left(1 - \frac{\alpha \cdot \left(1 - \|A\|_2\right)^2}{2n}\right) \cdot D(p^n \| q^n).$$
The proof uses the following "Converse" of Pinsker’s lemma

Let $q$ be a probability measure on a finite set $\mathcal{X}$, and assume:

$$\alpha \triangleq \min_x q(x) > 0.$$ 

Then:

$$D(p||q) \leq \frac{4}{\alpha} \cdot |p - q|^2_{TV}, \quad \forall p \in \mathbb{P}(\mathcal{X}).$$
Corollary

\[ D(p^n \Gamma^t || q^n) \to 0 \text{ exponentially, as } t \to \infty. \]
Distances between measures that are not Wasserstein distances

Example:

$\mathcal{X}^n$: product Borel space,
$p^n$ and $r^n$: Borel probability measures on $\mathcal{X}^n$.

\[ \tilde{W}(p^n, r^n) \triangleq \min_{\pi = \mathcal{L}(Y^n, Z^n) \in \Pi(p^n, r^n)} \sqrt{\sum_{i=1}^{n} Pr_\pi^2 \{Y_i \neq Z_i\}}. \]