Contractivity of Trace Distance

Formal statement

Let mathcal N:L(mathcal H(A))→ L(mathcal H(B)) be a quantum channel, meaning a completely positive trace-preserving linear map. For any two density operators ρ and σ on mathcal H(A), the trace distance cannot increase under mathcal N:

$$D(\mathcal{N}(\rho),\mathcal{N}(\sigma)) \leq D(\rho,\sigma),$$

where

$$D(\rho,\sigma) = \frac{1}{2}\|\rho-\sigma\|_1$$

and

$$\|X\|_1 = \operatorname{Tr}\sqrt{X^\dagger X}.$$

Equivalently,

$$\|\mathcal{N}(\rho)-\mathcal{N}(\sigma)\|_1 \leq \|\rho-\sigma\|_1.$$

This theorem is usually called the contractivity of trace distance, or the data-processing inequality for trace distance. The phrase “data processing” means that if two states are passed through the same physical process, their distinguishability cannot increase.

A slightly more general statement is also true. Complete positivity is more than is needed for this particular inequality. If Φ is positive and trace preserving, then

$$\|\Phi(X)\|_1 \leq \|X\|_1$$

for every Hermitian traceless operator X. Since X=ρ-σ is Hermitian and traceless, the trace-distance contraction follows. In quantum information, however, we usually state the theorem for quantum channels because completely positive trace-preserving maps are the physically valid deterministic operations on systems that may be entangled with references.

Operational meaning

Trace distance has a direct meaning in state discrimination. If ρ and σ are given with equal prior probabilities, the Helstrom theorem says that the optimal success probability for distinguishing them is

$$P_{\mathrm{succ}}^{\mathrm{opt}} = \frac{1}{2}\left(1 + D(\rho,\sigma)\right).$$

Thus D(ρ,σ) measures the best possible distinguishing advantage over random guessing.

The contractivity theorem says that applying the same quantum channel to both possible states cannot make them easier to distinguish. Noise, forgetting, coarse-graining, measurement without keeping all outcomes, partial trace, and decoherence may erase information. They cannot create new information about which state was originally prepared.

The clean mental image is this. Suppose Bob wants to distinguish ρ from σ. If the states first pass through a channel mathcal N, then Bob only sees mathcal N(ρ) or mathcal N(σ). Any measurement Bob performs after the channel could have been simulated before the channel by pulling the measurement backward through the channel. Therefore the best strategy after the channel cannot outperform the best strategy before the channel.

Proof using the adjoint map

We prove the theorem in finite dimensions. Let

$$X = \rho - \sigma.$$

Then X is Hermitian and traceless. The trace norm of a Hermitian operator has the variational characterization

$$\|X\|_1 = \max_{-I \le H \le I} \mathrm{Tr}(HX),$$

where the maximum is over Hermitian operators H satisfying -I≤ H≤ I.

Let mathcal N^† denote the Hilbert-Schmidt adjoint of mathcal N, defined by

$$\mathrm{Tr}[Y\,\mathcal{N}(X)] = \mathrm{Tr}\big[\mathcal{N}^\dagger(Y)\,X\big]$$

for all operators X and Y of compatible dimensions. Since mathcal N is trace preserving, its adjoint is unital:

$$\mathcal{N}^\dagger(I_B) = I_A.$$

Since mathcal N is positive, its adjoint is also positive. Therefore, whenever

$$-I_B \le H \le I_B,$$

we have

$$0 \le I_B + H \le 2I_B$$

and

$$0 \le I_B - H \le 2I_B.$$

Applying the positive unital map mathcal N^†, we get

$$0 \le I_A + \mathcal{N}^\dagger(H) \le 2I_A$$

and

$$0 \le I_A - \mathcal{N}^\dagger(H) \le 2I_A.$$

Hence

$$-I_A \le \mathcal{N}^\dagger(H) \le I_A.$$

Now compute:

$$\begin{aligned}\|\mathcal{N}(X)\|_1 &= \max_{-I_B \le H \le I_B} \mathrm{Tr}[H\,\mathcal{N}(X)] \\ &= \max_{-I_B \le H \le I_B} \mathrm{Tr}\left[\mathcal{N}^\dagger(H)X\right].\end{aligned}$$

But every operator mathcal N^†(H) appearing in this maximum is a valid test operator in the variational formula for \|X\|₁, because

$$-I_A \le \mathcal{N}^\dagger(H) \le I_A.$$

Therefore

$$\|\mathcal{N}(X)\|_1 \le \max_{-I_A \le K \le I_A} \mathrm{Tr}(KX) = \|X\|_1.$$

Substituting X=ρ-σ, we obtain

$$\|\mathcal{N}(\rho) - \mathcal{N}(\sigma)\|_1 \le \|\rho - \sigma\|_1.$$

Dividing by 2 gives

$$D(\mathcal{N}(\rho), \mathcal{N}(\sigma)) \le D(\rho, \sigma).$$

This proves the theorem.

Proof using measurements

The same theorem has an even more operational proof. Let Bob perform a POVM M₍y₎\ on the output system B. If the input was ρ, the outcome distribution is

$$p_y = \mathrm{Tr}(M_y \mathcal{N}(\rho)).$$

If the input was σ, the outcome distribution is

$$q_y = \mathrm{Tr}(M_y \mathcal{N}(\sigma)).$$

Using the adjoint channel,

$$p_y = \mathrm{Tr}(\mathcal{N}^\dagger(M_y)\rho), \qquad q_y = \mathrm{Tr}(\mathcal{N}^\dagger(M_y)\sigma).$$

Because mathcal N^† is positive and unital, the operators

$$\widetilde{M}_y = \mathcal{N}^\dagger(M_y)$$

form a POVM on the input system:

$$\widetilde{M}_y \geq 0, \qquad \sum_y \widetilde{M}_y = \mathcal{N}^\dagger\left(\sum_y M_y\right) = \mathcal{N}^\dagger(I_B) = I_A.$$

Therefore every measurement after the channel is equivalent to some measurement before the channel. Since the trace distance is the maximum classical total variation distance over all measurements,

$$D(\rho,\sigma) = \max_{\{M_y\}} \frac{1}{2} \sum_y \left|\mathrm{Tr}(M_y\rho) - \mathrm{Tr}(M_y\sigma)\right|,$$

the best distinguishability after the channel cannot exceed the best distinguishability before the channel.

This proof explains the theorem in one sentence: processing the states first only restricts the measurements available later; it cannot give a measurement advantage that was not already available before the processing.

Example: unitary evolution preserves trace distance

Let

$$\mathcal{U}(\rho) = U\rho U^\dagger$$

for a unitary U. Then

$$\mathcal{U}(\rho) - \mathcal{U}(\sigma) = U(\rho - \sigma)U^\dagger.$$

The trace norm is invariant under unitary conjugation:

$$\|UXU^\dagger\|_1 = \|X\|_1.$$

Therefore

$$D(\mathcal{U}(\rho),\mathcal{U}(\sigma)) = D(\rho,\sigma).$$

This is the reversible case. A unitary channel does not lose information. If two states were distinguishable before a unitary transformation, they are equally distinguishable after it. The optimal measurement may rotate, but the amount of distinguishability remains unchanged.

Example: complete dephasing can destroy distinguishability

Consider the complete dephasing channel in the computational basis:

$$\Delta_Z(\rho) = |0\rangle\langle 0|\rho|0\rangle\langle 0| + |1\rangle\langle 1|\rho|1\rangle\langle 1|.$$

Now compare the two pure states

$$|+\rangle = \frac{|0\rangle + |1\rangle}{\sqrt{2}}, \qquad |-\rangle = \frac{|0\rangle - |1\rangle}{\sqrt{2}}.$$

Before dephasing, these states are orthogonal, so

$$D(|+\rangle\langle +|, |-\rangle\langle -|) = 1.$$

After dephasing,

$$\Delta_Z(|+\rangle\langle +|) = \frac{I}{2},$$

and

$$\Delta_Z(|-\rangle\langle -|) = \frac{I}{2}.$$

Thus

$$D(\Delta_Z(|+\rangle\langle+|), \Delta_Z(|-\rangle\langle-|)) = D(I/2, I/2) = 0.$$

This is the simplest operational picture of decoherence. The two states were perfectly distinguishable because they differed by phase coherence. The dephasing channel erased precisely that coherence, making them identical.

Example: partial trace removes information

Consider the two Bell states

$$|\Phi^+\rangle_{AB} = \frac{|00\rangle + |11\rangle}{\sqrt{2}},$$

and

$$|\Phi^-\rangle_{AB} = \frac{|00\rangle - |11\rangle}{\sqrt{2}}.$$

They are orthogonal, so globally

$$D(|\Phi^+\rangle\langle\Phi^+|, |\Phi^-\rangle\langle\Phi^-|) = 1.$$

Now apply the channel that discards system B:

$$\mathcal{N}(\tau_{AB}) = \mathrm{Tr}_B(\tau_{AB}).$$

Both reduced states on A are maximally mixed:

$$\mathrm{Tr}_B(|\Phi^+\rangle\langle\Phi^+|) = \frac{I}{2},$$

and

$$\mathrm{Tr}_B(|\Phi^-\rangle\langle\Phi^-|) = \frac{I}{2}.$$

Therefore

$$D\big(\mathrm{Tr}_B\,|\Phi^+\rangle\langle\Phi^+|, \mathrm{Tr}_B\,|\Phi^-\rangle\langle\Phi^-|\big) = 0.$$

The sign difference between the two Bell states is stored in a global correlation. If we discard B, system A alone contains no information about that sign. Contractivity says exactly that forgetting a subsystem cannot improve distinguishability.

Example: depolarizing noise shrinks the Bloch vector

For a qubit, define the depolarizing channel

$$\mathcal{D}_p(\rho) = (1-p)\rho + p\frac{I}{2}, \qquad 0 \le p \le 1.$$

For any two qubit states ρ and σ,

$$\mathcal{D}_p(\rho) - \mathcal{D}_p(\sigma) = (1-p)(\rho - \sigma).$$

Therefore

$$D(\mathcal{D}_p(\rho), \mathcal{D}_p(\sigma)) = (1-p)D(\rho, \sigma).$$

This example shows a clean quantitative contraction. Depolarizing noise pulls every state toward the center of the Bloch ball. Differences between states shrink by the factor 1-p. When p=0, the channel is the identity and trace distance is preserved. When p=1, all inputs become I/2, and every pair of outputs has trace distance zero.

Example: classical stochastic maps

The theorem also contains the classical statement that stochastic processing cannot increase total variation distance. Suppose

$$\rho = \sum_x p_x |x\rangle\langle x|, \qquad \sigma = \sum_x q_x |x\rangle\langle x|$$

are diagonal classical states. Their trace distance is the total variation distance:

$$D(\rho, \sigma) = \frac{1}{2} \sum_x |p_x - q_x|.$$

A classical noisy channel is a stochastic matrix T(y|x). It maps

$$p_x \mapsto p'_y = \sum_x T(y|x) p_x, \qquad q_x \mapsto q'_y = \sum_x T(y|x) q_x.$$

The quantum contractivity theorem reduces to

$$\frac{1}{2} \sum_y |p'_y - q'_y| \leq \frac{1}{2} \sum_x |p_x - q_x|.$$

So the quantum theorem is a noncommutative generalization of an ordinary fact from classical probability: post-processing data cannot increase statistical distinguishability.

Why equality sometimes holds and sometimes fails

Equality holds for reversible channels, such as unitary channels. It can also hold for a particular pair of states under a nonunitary channel if the channel preserves exactly the information needed to distinguish that pair. For example, complete dephasing preserves the trace distance between computational-basis states $|0\rangle$ and $|1\rangle$, because those states are already diagonal in the dephasing basis.

But equality fails whenever the channel erases information relevant to distinguishing the states. Dephasing erases phase information. Partial trace erases correlations with discarded systems. Depolarizing noise shrinks all Bloch-vector differences. In these cases, the trace distance strictly decreases.

This is why contractivity is not merely a mathematical inequality. It is a statement about information flow. A decrease in trace distance means that some information distinguishing the two alternatives has been lost to an inaccessible system, randomized away, or coarse-grained.

Relation to the data-processing principle

The contractivity theorem is a special case of the broader data-processing principle in quantum information theory. Many distinguishability measures decrease under channels. Quantum relative entropy satisfies

$$D(\rho\|\sigma) \geq D(\mathcal{N}(\rho)\|\mathcal{N}(\sigma)),$$

and fidelity satisfies the opposite monotonicity direction,

$$F(\mathcal{N}(\rho),\mathcal{N}(\sigma)) \geq F(\rho,\sigma).$$

Trace distance follows the distinguishability convention: larger means easier to distinguish, so it cannot increase under processing. Fidelity follows the closeness convention: larger means closer, so it cannot decrease under processing.

Together with the Fuchs-van de Graaf inequalities, these facts explain why trace distance and fidelity are the two most widely used state-comparison tools. Trace distance gives direct operational distinguishability. Fidelity gives geometric overlap and purification-based closeness. Contractivity says that physical processing respects both notions.

How to use the theorem

In practice, the theorem is often used to avoid calculations. If a complicated operation mathcal N is applied to two states, one immediately has

$$D(\mathcal{N}(\rho),\mathcal{N}(\sigma)) \leq D(\rho,\sigma).$$

For example, if

$$D(\rho,\sigma) \leq \varepsilon,$$

then after any channel,

$$D(\mathcal{N}(\rho),\mathcal{N}(\sigma)) \leq \varepsilon.$$

Thus trace-distance security, correctness, or approximation guarantees are stable under all later quantum processing.

This is especially important in cryptography. If the real state and ideal state are close in trace distance, then no adversary can make them more distinguishable by applying a quantum operation. Any attack, measurement, storage process, or post-processing map is itself a channel. Therefore the trace-distance guarantee remains valid after the adversary acts.

It is also crucial in error analysis. If two implementations prepare states within trace distance ε, then any subsequent circuit, measurement, or noise process cannot increase that state-preparation error as measured by trace distance.

Common mistakes

A common mistake is to think that a channel can make two states more distinguishable by “amplifying” their difference. Deterministic quantum channels cannot do this when both states pass through the same channel. A physical process may make some features more visible, but it cannot increase the optimal distinguishability beyond what was already present in the input states.

A second mistake is to forget that the same channel must be applied to both states. If two different operations are applied depending on which state was prepared, then the theorem does not apply; that would already insert information about the hypothesis into the dynamics.

A third mistake is to confuse trace distance with Hilbert-Schmidt distance. The trace distance is contractive under quantum channels. The Hilbert-Schmidt distance

$$\|\rho - \sigma\|_2$$

is not generally contractive under all quantum channels. This is one reason trace distance, rather than Hilbert-Schmidt distance, has a privileged operational role.

A fourth mistake is to think that contractivity means all channels strictly reduce distinguishability. Unitary channels preserve trace distance exactly, and nonunitary channels may preserve trace distance for special pairs of states. The theorem says “cannot increase,” not “must decrease.”

Final mental image

Trace distance measures how distinguishable two states are by the best possible measurement. A quantum channel is a physical processing step applied equally to both states. Any measurement after the channel can be simulated by a corresponding measurement before the channel. Therefore the channel cannot improve the best possible discrimination strategy.

In one sentence:

physical processing cannot create distinguishability between states.

Mathematically,

$$D(\mathcal{N}(\rho),\mathcal{N}(\sigma)) \leq D(\rho,\sigma).$$

Operationally, this is the statement that noise, forgetting, coarse-graining, and open-system evolution cannot make two unknown quantum states easier to tell apart when the same process acts on both.

References

Helstrom, Carl W. Quantum Detection and Estimation Theory. Academic Press, 1976.

Nielsen, Michael A., and Isaac L. Chuang. Quantum Computation and Quantum Information. Cambridge University Press, 10th anniversary edition, 2010. See the discussion of trace distance and its monotonicity under quantum operations.

Watrous, John. The Theory of Quantum Information. Cambridge University Press, 2018. See the sections on trace norm, trace distance, measurements, and distinguishability.

Fuchs, Christopher A., and Jeroen van de Graaf. “Cryptographic Distinguishability Measures for Quantum-Mechanical States.” IEEE Transactions on Information Theory 45, no. 4 (1999): 1216–1227.

Pérez-García, David, Michael M. Wolf, Dénes Petz, and Mary Beth Ruskai. “Contractivity of Positive and Trace-Preserving Maps under L_p-Norms.” Journal of Mathematical Physics 47, 083506 (2006).

Wilde, Mark M. Quantum Information Theory. Cambridge University Press, 2nd edition, 2017.

Document information

Contractivity of Trace Distance

Project	[QIT 002] State Distinguishability and Measurement Theorems
Document	Primary document
Author	mujirin
Verifier	Not verified
Downloaded	July 04, 2026 23:31 KST
Status	Working
Document link	https://www.theorytrace.com/projects/state-distinguishability-and-measurement-theorems/documents/untitled-document-ddee6b/