Generalizing the Posterior Matching Scheme to Higher Dimensions via Optimal Transportation

Rui Ma and Todd P. Coleman
Department of Bioengineering
University of California, San Diego
La Jolla, CA

***Abstract*—This paper re-visits Shayevitz & Feder's recent 'Posterior Matching Scheme', an explicit, dynamical system encoder for communication with feedback that treats the message as a point on the $[0,1]$ line and achieves capacity on memoryless channels. It has two key properties that ensure that it maximizes mutual information at each step: (a) the encoder sequentially hands the decoder what is missing; and (b) the next input has the desired statistics. Motivated by brain-machine interface applications and multi-antenna communications, we consider developing dynamical system feedback encoders for scenarios when the message point lies in higher dimensions. We develop a necessary and sufficient condition - the Jacobian equation - for any dynamical system encoder that maximizes mutual information. In general, there are many solutions to this equation. We connect this to the Monge-Kantorovich Optimal Transportation Problem, which provides a framework to identify a unique solution suiting a specific purpose. We provide two examplar capacity-achieving solutions - for different purposes - for the multi-antenna Gaussian channel with feedback. This insight further elucidates an interesting relationship between interactive decision theory problems and the theory of optimal transportation.**

I. Introduction

Consider a communication channel $P_{Y|X}$ with feedback as shown in figure 1 where the objective is to transmit a message W with feedback. A causal encoder at time i specifies $X_i = e_i(W, Y^{i-1})$. A rate R is "achievable" if for $n \gg 1$ and $|\mathsf{W}| = 2^{nR}$, the posterior distribution becomes a point mass at W. Given the channel $P_{Y|X}$ and a cost function $\eta : \mathsf{X} \to \mathbb{R}_+$, the capacity-cost function is given by

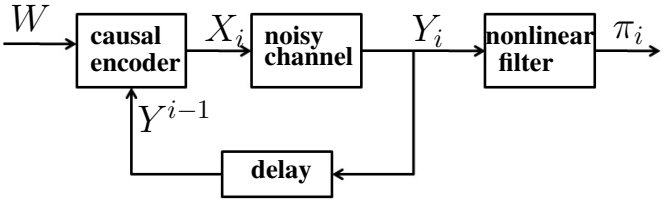

Fig. 1. Communication of a message point W with causal feedback over a memoryless channel.

$$C\left(\eta, P_{Y|X}, L\right) \triangleq \max_{P_X : \mathbb{E}[\eta(X)] \leq L} I(P_X, P_{Y|X}). \quad (1)$$

where $I(P_X, P_{Y|X})$ is the mutual information over the channel when the input X has statistics P_X. It is known that $C\left(\eta, P_{Y|X}, L\right)$ is the fundamental limit of communication over channel $P_{Y|X}$ over all encoders that whose average attain cost $\eta(X)$ is upper-bounded L [1].

Recent developments in the information theoretic literature [2], [3], [4] have espoused using an alternative 'analog message point' viewpoint where $\mathsf{W} = [0,1]$, $\mathsf{X} \subset \mathbb{R}$, the notion of a rate R being "achievable" is cast in terms of the speed of convergence of the decoder's posterior probability distribution towards a point mass at W. At each time step (as compared to only at time n), the decoder specifies a posterior belief π_i about the message given $Y_1, \ldots, Y_i$: $\pi_i(A) \equiv \mathbb{P}\left(W_i \in A | Y^i\right)$. A rate R is achievable in this analog scenario if

$$\pi_n\left(\{u : Q_{nR}(u) = Q_{nR}(W)\}\right) \to 1$$

as $n \to \infty$, where $Q_{nR} : [0,1] \to \{1, \ldots, 2^{nR}\}$ is a sequence of uniform quantizers. The fundamental limits of this problem and the standard block coding viewpoint with $|\mathsf{W}| = 2^{nR}$ are equivalent [2].

We note that the following standard Lemma from information theory [5]:

Lemma 1.1: Fix an $n \geq 1$. If a feedback encoder $\left(X_i = e_i(W, Y^{i-1}) : i = 1, \ldots, n\right)$ satisfies the constraint

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\eta(X_i)\right] \leq L, \quad (2)$$

then

$$\frac{1}{n} I(W; Y^n) \leq C\left(\eta, P_{Y|X}, L\right). \quad (3)$$

Equality holds if and only if

(a) $Y_1 \ldots, Y_n$ are statistically independent.
(b) $X_i \sim P_X^*$ for each i, where P_X^* is the capacity-achieving distribution in (1).

If we define F_X as the cumulative distribution function (CDF) pertaining to P_X^*, then the posterior matching (PM) scheme [2] is defined as follows:

$$W_1 = W, \qquad W_{i+1} = F_{W|Y^i}(W|Y^i), \quad (4)$$

$$X_i = F_X^{-1}(W_i). \quad (5)$$

The PM scheme extends previously developed continuous message point schemes [6], [7] to arbitrary memoryless channels when $\mathsf{X} \subset \mathbb{R}$. It is the optimal solution [3] to a team decision theory problem between the encoder and decoder where the objective is to maximize mutual information $I(W; Y^n)$. Under mild assumptions, it achieves capacity [2].

A. Time-Invariant Representation

An equivalent representation to (4) can be given as follows [2]:

$$W_1 = W, \qquad W_{i+1} = F_{W_i|Y_i}(W_i|Y_i) \equiv S_{Y_i}(W_i) \tag{6}$$

This viewpoint is pleasing computationally because there is no dependence on the time i, i.e. $\{S_y : y \in \mathsf{Y}\}$ is a *fixed, time-invariant* set of maps. The essence of the scheme can be boiled down to the following:

(a) The encoder sequentially hands the decoder what is missing.
(b) The next input has the desired statistics.

To elucidate (a) in more depth, (4)-(5) demonstrates that X_{i+1} is statistically independent of Y^i; Because the channel is memoryless, it follows that Y_{i+1} is statistically independent of Y^i. To elucidate (b) in more depth, (4) shows that W_{i+1} is uniform on $[0, 1]$ and so from (5), each $X_{i+1} \sim P_X$. Properties (a) and (b) guarantee [2] that for **any** n, (3) holds with equality: $\frac{1}{n} I(W; Y^n) = C\left(\eta, P_{Y|X}, L\right)$. Moreover, under mild technical conditions, the PM scheme also achieves reliability at any rate $R < C$ [2].

The 'continuous message point' problem formulation for reliable communication is pleasing for applications beyond traditional digital communications - such as biological communication, network control [8], and brain-machine interfaces [9], the uncertain message is fundamentally in a continuum, because

- the alphabet of the message, W, is fixed *irrespective* of the number n of uses of the channel or any notion of rate of communication.
- A 'rate' R being achievable is defined in terms of whether or not the decoder's posterior belief converges quickly enough to a point mass.

The PM scheme, which is optimal in this analog message point scenario, is desirable for implementation because

- There is no forward error correction - it simply adapts on the fly and sequentially hands the decoder what is missing.
- The scheme admits a simple time-invariant dynamical system structure.

These properties of the PM scheme have made it amenable to implementation in real-world systems coupling computers with physical/biological systems that practically achieve fundamental limits [8],[9].

B. Motivation and Main Results

For message point communication paradigms as discussed above with $\mathsf{W} = [0, 1]$, the key properties required so that this paradigm has the same fundamental limits as the standard Shannon-theoretic approach are:

- W should be *uncountable*, so that an increasingly finer set of quantizers $(Q_n : \mathsf{W} \to \{1, \ldots, 2^{nR}\} : \ n \geq 1)$ can be described,
- W should be *compact*, so that every open cover (pertaining to quantization intervals) has a finite subcover (in particular, at time n, there are 2^{nR} of them).

As such, one naturally asks the question of whether or not it is desirable to consider W to be a more general uncountable, compact set. When the channel input space $\mathsf{X} \subset \mathbb{R}$, the PM scheme in (4)-(6) suffices. But in other scenarios, $\dim(\mathsf{X}) = d$, for example $\mathsf{X} = \mathbb{R}^d$. In such scenarios, perhaps it is more natural to consider paradigms where $\dim(\mathsf{W}) \geq \dim(\mathsf{X})$ so that the mapping ϕ pertaining $X_{i+1} = \phi(W_{i+1})$ has simple structure.

Suppose, for example, if X_i pertains to the input of a multi-antenna communication channel and so $X_i \in \mathbb{R}^d$. Another example is the relationship between feedback information theory and control over noisy channels. In characterizing the fundamental limits of such problems, it is of paramount importance to consider feedback communication where the $\dim(\mathsf{W}) = \dim(\mathsf{X})$ [8, Thm 4.3]. It is unclear a priori how to directly generalize (4)-(6) using cumulative distribution functions - because their outputs are on the $[0, 1]$ line.

In this paper, we attempt to generalize the PM scheme to compact, uncountable alphabets W where $\mathsf{W} \subset \mathbb{R}^d$ for some $d \geq 1$. In Section III, we develop a first-order differential equation - termed the Jacobian equation - that provides necessary and sufficient conditions on maps S_y that guarantee that $I(W; Y^n) = nC$. We show how when $\mathsf{W} = [0, 1]$, the posterior matching scheme satisfies this equation. In Section IV, we connect the aforementioned Jacobian equation to the theory of optimal transportation (OTT) [10]. With different OTT cost functions, we show that different solutions of the Jacobian equation minimizes cost. For the Euclidean squared cost function, we show that a second-order equation characterizes a unique OTT-optimal map, termed the Brenier coupling. For a different weighted Euclidean cost function, we show that the Knothe-Rosenblatt coupling [11] is optimal. Under appropriate technical conditions, an OTT-optimal scheme attains reliability. However, the different cost functions provide different ways that reliability is attained - for example,the Brenier coupling enables convergence of each dimension of the message to be equally as important. On the flipside, the Knothe-Rosenblatt coupling inherently prefers one dimension to converge and to then 'onion-peel' for each subsequent dimension. As such, the Knothe-Rosenblatt coupling enables a notion of 'unequal error protection'. We demonstrate an example of the differences between these couplings with a multi-antenna Gaussian channel with Note that optimal transportation theory was recently used [12] to elucidate properties of the optimal strategy for Witsenhausen's counterexample—another team decision theory problem involving interaction between two agents. As such, those results along with these point to optimal transportation theory being a very useful lens to understand interactive team decision theory problems.

II. Preliminaries and Definitions

- For a sequence $a_1, a_2, \ldots$, denote $a^j \triangleq (a_1, \ldots, a_j)$.

- We denote $U \perp\!\!\!\perp V$ to mean that the random variables U and V are statistically independent.
- For an input distribution P_X and a noisy channel with conditional distribution $P_{Y|X}$, denote $I(P_X, P_{Y|X})$ as the mutual information:

$$I(P_X, P_{Y|X}) = \int_{\mathsf{X}\times\mathsf{Y}} \log \frac{dP_{Y|X=x}}{dP_Y}(y) P_{Y|X=x}(dy) P_X(dx)$$

- For a Borel space W, denote the set of probability measures on W as $\mathcal{P}(\mathsf{W})$.
- We define the set of couplings $\mathcal{C}(\mathbb{P}, \mathbb{Q})$ to be the set of joint distributions $P_{U,V}$ for which $P_U = \mathbb{P}$ and $P_V = \mathbb{Q}$.
- Given two probability measure $\mathbb{P}, \mathbb{Q}$ defined on measurable space W and a Borel-measurable map $S : \mathsf{W} \to \mathsf{W}$, we say that S *pushes* $\mathbb{P}$ forward to $\mathbb{Q}$, specified as $S\#\mathbb{P} = \mathbb{Q}$, if a random variable W with distribution $\mathbb{P}$ and map S results in random variable $V = S(W)$ having distribution $\mathbb{Q}$.
- Denote the set of all maps that push $\mathbb{P}$ to $\mathbb{Q}$ as $\mathsf{S}(\mathbb{P}, \mathbb{Q})$:

$$\mathsf{S}(\mathbb{P}, \mathbb{Q}) \triangleq \{S : \mathsf{W} \to \mathsf{W} \; s.t. \; S\#\mathbb{P} = \mathbb{Q}\}. \tag{7}$$

- A map $S : \mathsf{W} \to \mathsf{W}$ is a *diffeomorphism* if S is invertible, S is differentiable, and S^{-1} is differentiable.
- With this, we have the following Lemma:
  Lemma 2.1: Consider two probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on a Borel space $\mathsf{W} \subset \mathbb{R}^d$ that have probability density functions $p(w)$ and $q(w)$ respectively with respect to the Lebesgue measure. Then the set of all diffeomorphisms in $\mathsf{S}(\mathbb{P}, \mathbb{Q})$ is the set of all maps $S : \mathsf{W} \to \mathsf{W}$ for which

$$p(w) = q(S(w))|J_S(S(w))|, \tag{8}$$

  where $|\cdot|$ denotes the matrix determinant and J_S is the Jacobian operator on the map S.
  For example, if $d = 2$, $w = (w[1], w[2])$, and $S = (S^1(\cdot), S^2(\cdot))$, then:

$$J_S(w) = \begin{bmatrix} \frac{\partial}{\partial w[1]} S^1(w) & \frac{\partial}{\partial w[2]} S^1(w) \\ \frac{\partial}{\partial w[1]} S^2(w) & \frac{\partial}{\partial w[2]} S^2(w) \end{bmatrix}$$

  Note that (8) is simply a consequence of the Jacobian relation in basic probability theory for invertible differentiable transformations of continuous random vectors. This is called the *Jacobian equation*.

III. The Jacobian Equation for Posterior Matching in Arbitrary Dimension

In this section, we consider the message space, W, to be an arbitrary compact, uncountable subset of $\mathbb{R}^d$. We define the notion of PM-compatibility for which the encoder is "handing the decoder what is missing" and $I(W; Y^n) = nC$ for any $n \geq 1$.

Consider the triple $(P_{Y|X}, \eta, L)$. Then $C\left(\eta, P_{Y|X}, L\right)$ is defined according to (1). Define the capacity-achieving input distribution P_X^* to be the distribution for which $I(P_X^*, P_{Y|X}) = C\left(\eta, P_{Y|X}, L\right)$. Assume that $\dim(\mathsf{W}) = \dim(\mathsf{X})$. Then there exists a distribution P_W and $\phi : \mathsf{W} \to \mathsf{X}$ for which $X = \phi(W)$ is distributed according to P_X^* when $W \sim P_W$. Since W is a compact, uncountable subset of $\mathbb{R}^d$, without loss of generality we assume that P_W has a density with respect to the Lebesgue measure. Consider the set of maps $(S_y : \mathsf{W} \to \mathsf{W})_{y\in\mathsf{Y}}$ and the map $\phi : \mathsf{W} \to \mathsf{X}$.

Definition 3.1: Given a message point distribution P_W, a map $\phi : \mathsf{W} \to \mathsf{X}$, and a noisy channel $P_{Y|X}$, we say that $(S_y)_{y\in\mathsf{Y}}$ is *PM-compatible* if the encoder scheme with time-invariant dynamics given by

$$W_1 = W, \qquad W_{i+1} = S_{Y_i}(W_i) \tag{9a}$$

$$X_{i+1} = \phi(W_{i+1}) \tag{9b}$$

satisfies

- (i) $S_y \# P_{W_1|Y_1=y} = P_W$ for all $y \in \mathsf{Y}$.
- (ii) S_y is a diffeomorphism for all $y \in \mathsf{Y}$.

The intuition behind the definition can be done for $n = 2$ and as we shall see, it generalizes for arbitrary n. Assuming $n = 2$, the idea is that, given $Y_1 = y$, we have one distribution $\mathbb{P} \equiv P_{W_1|Y_1=y}$. At time step 2, the encoder has this distribution and it would like to apply the map $W_2 = S_y(W_1)$. Note that, as a consequence, S_y pushes $\mathbb{P}$ forward to $\mathbb{Q} \equiv P_{W_2|Y_1=y}$. In order to maximize mutual information, from condition (a) in Lemma 1.1, we need to guarantee that Y_1, Y_2 are statistically independent. Because the channel is memoryless, this is the same as making W_2 statistically independent of Y_1. As such, we desire that $\mathbb{Q} \equiv P_{W_2|Y_1=y} = P_{W_2}$. To satisfy condition (b) in Lemma 1.1, it is our desire to make $P_{W_2} = P_W$, where $\phi \# P_W = P_X^*$. Combining these together, we arrive at the aforementioned definition.

Corollary 3.2: If ϕ is invertible, then "PM-compatibility" is equivalent to replacing (9) with

$$X_1 = \phi(W), \qquad X_{i+1} = \phi \circ S_{Y_i} \circ \phi^{-1}(X_i) \tag{10}$$

and replacing (i) with $(\phi \circ S_y \circ \phi^{-1}) \# P_{X_1|Y_1=y} = P_X$.

With this, we can now characterize necessary and sufficient conditions for PM-compatibility in terms of the first-order Jacobian equation in (8):

Lemma 3.3: The set of diffeomorphisms $(S_y)_{y\in\mathsf{Y}}$ is PM-compatible if and only if

$$f_{W_1|Y_1=y}(w) = f_W\left(S_y(w)\right)|J_{S_y}(w)|. \tag{11}$$

Proof: Invoke Definition 3.1 and Lemma 2.1 with $\mathbb{P} \equiv P_{W_1|Y_1=y}$ and $\mathbb{Q} \equiv P_{W_2|Y_1=y} = P_W$. ∎

Remark 1: If $\mathsf{W} = [0, 1]$ and $f_W(w) = 1$, we now show how we can recover the posterior matching scheme in [2]. If we constraint $S_y : [0, 1] \to [0, 1]$ to be an increasing function, then (11) becomes

$$S_y'(w) = f_{W|Y_1=y}(w). \tag{12}$$

If we impose the boundary condition that $S_y(0) = 0$, then we recover

$$S_y(w) = F_{W|Y_1=y}(w) \tag{13}$$

as a solution to the Jacobian equation. Note that this is precisely the 'Posterior Matching' scheme developed by

Shayevitz & Feder [2]. Analogously, if we constrain S_y to be decreasing with $S_y(0) = 1$, then we see that $S_y(w) = 1 - F_{W|Y_1=y}(w)$ is also a solution to the Jacobian equation and thus PM-compatible. This means that there are many maps that are PM-compatible.

With this, we can state the following Lemma that shows why PM-compatibility is fundamentally related to maximizing mutual information for arbitrary block length with analog message points:

Lemma 3.4: If $(S_y)_{y\in\mathsf{Y}}$ is PM-compatible, then for an encoder given by (9) and any $n \geq 1$:

- (a) W can be recovered from W_{n+1} and Y^n
- (b) $\frac{1}{n}I(W;Y^n) = I(P_X, P_{Y|X})$ where P_X is the induced distribution on the random variable $X = \phi(W)$ with $W \sim P_W$.

Proof: We first prove (a), which shows in another sense that the encoder is 'handing the decoder what is missing'. Since S_y is a diffeomorphism, it is invertible. As such, apply the following recursively:

$$W_i = S_{Y_i}^{-1}(W_{i+1}). \tag{14}$$

from $i = n$ downward to $i = 1$.

To prove (b), we do a proof by induction. Clearly when $i = 2$, it follows that $W_2 \perp\!\!\!\perp Y_1$, which implies that $Y_2 \perp\!\!\!\perp Y_1$ since the channel is memoryless. Moreover, since $P_{W_2} = P_{W_1}$, it follows that (Y_1, Y_2) are i.i.d.. Suppose by induction that $W_k \perp\!\!\!\perp Y^{k-1}$. Then we have that the pair $(W_k, Y_k) \perp\!\!\!\perp Y^{k-1}$ because the channel is memoryless. Therefore:

$$\begin{aligned} P_{W_k,Y_k|Y^{k-1}} &= P_{Y_k|Y^{k-1}} P_{W_k|Y^k} = P_{Y_k} P_{W_k|Y^k} && (15)\\ P_{W_k,Y_k|Y^{k-1}} &= P_{W_k,Y_k} = P_{Y_k} P_{W_k|Y_k} && (16)\\ \Rightarrow P_{W_k|Y^k} &= P_{W_k|Y_k}. && (17) \end{aligned}$$

where the latter equality in (15) and the former equality in (16) both follow because $(W_k, Y_k) \perp\!\!\!\perp Y^{k-1}$. Therefore,

$$\begin{aligned} & f_{W_{k+1}|Y^k=y^k}(w_{k+1}) \\ =\ & f_{W_k|Y^k=y^k}\left(S_{y_k}^{-1}(w_{k+1})\right) \frac{1}{|J_{S_{y_k}}(S_{y_k}^{-1}(w_{k+1}))|} && (18)\\ =\ & f_{W_k|Y_k=y_k}\left(S_{y_k}^{-1}(w_{k+1})\right) \frac{1}{|J_{S_{y_k}}(S_{y_k}^{-1}(w_{k+1}))|} && (19)\\ =\ & f_W(w_{k+1}) && (20) \end{aligned}$$

where (18) follows from Lemma 2.1; (19) follows from (17); and (20) follows from (11). ∎

Note that Definition 3.1 (i) guarantees that "the decoder is giving what the encoder is missing" because then $I(W;Y^n) = nI(P_X, Q_{Y|X})$ is maximal, and also one can recover W from W_{i+1} and Y^i. The invertibility of S_y in property (ii) is crucially important so that the necessary condition $\frac{1}{n}I(W;Y^n) = I(P_X, Q)$ pertaining to reliable communication of any rate $R < C$ holds, but also for sufficiency (see [2] and [4, Lemma 6.1(iv)]). The differentiability of S_y in property (ii) of Definition 3.1 is stated so that we can guarantee that a density exists for the distribution $P_{W|Y=y}$, which we want to be $P_W = \text{unif}[0,1]$.

IV. Connection to Optimal Transportation Theory

In this section, we now consider a way to find desirable PM-compatible schemes by associating a cost function whose expectation with respect to a coupling gives an average cost. The optimal transportation problem (OT) is the problem of finding an optimal coupling - which under mild assumptions is unique. Monge was the first to formulate this problem [13], while Kantorovich reformulated the problem in a more general sense and made significant contribution to the theory [14].

The physical interpretation of the OT problem is to imagine moving one fixed pile of dirt underground from one location to form a new pile of dirt above ground at another location, such that the new pile has the same mass and conforms to specific constraints on its shape. The goal is to find a moving plan which minimizes the total cost incurred (i.e. fuel consumption) under a certain cost function, while conforming to the shape constraints. Because mass is conserved, we can abstract both piles as probability measures $\mathbb{P}$ and $\mathbb{Q}$. See Figure 2.

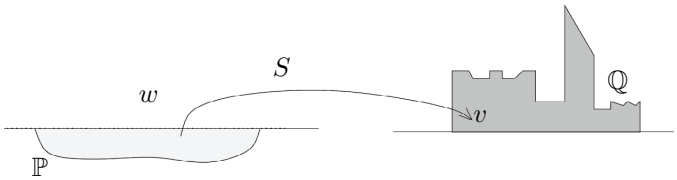

Fig. 2. The map S that moves a fixed pile of underground dirt from one location to a pile of dirt above ground at another location with a desired shape. Both piles can be interpreted as probability measures $\mathbb{P}$ and $\mathbb{Q}$.

With this, we can formally define the optimal transportation problem.

Definition 4.1 (Optimal Transportation Problem): For any $\mathbb{P} \in \mathcal{P}(\mathsf{W}), \mathbb{Q} \in \mathcal{P}(\mathsf{W})$, and $c : \mathsf{W} \times \mathsf{W} \to [0, \infty]$, define the optimal transportation cost as

$$L(\mathbb{P}, \mathbb{Q}, c) = \inf_{P_{U,V} \in \mathcal{C}(\mathbb{P},\mathbb{Q})} \int_{\mathsf{W}\times\mathsf{W}} c(u,v) P_{U,V}(du, dv) \tag{21}$$

Any $P^*_{U,V}$ that attains this infimum is termed an 'optimal coupling'.

This problem has been studied in depth in [15], [10], [16].

We now have the following general theorem that was presented by Kontarovich

Theorem 4.2: Suppose $\mathsf{W} \subset \mathbb{R}^d$, $\mathbb{P}$ and $\mathbb{Q}$ induce probability density functions $p(w)$ and $q(w)$ respectively with respect to the Lebesgue measure. If the cost function $c(w,v) = h(w-v)$ where h is strictly convex, then there exists a unique optimal coupling $P^*_{U,V}$ that is given by a diffeomorphism S for which $V = S(W)$ and S satisfies the Jacobian equation

$$p(w) = q(S(w)) \cdot |\det J_S(S(w))|, \tag{22}$$

Note that (22) follows from (8). What is interesting about the above theorem is the uniqueness of the optimal solution. See [15] for proof of the above theorem. As such, we have the following theorem:

Theorem 4.3 (Generalized Posterior-Matching Theorem): Let $\mathbb{P} = P_{W_1|Y_1=y} \equiv P_{W|Y=y}$, $\mathbb{Q} = P_W$, and $c : \mathsf{W} \to \mathsf{W}$

satisfy $c(w,v) = h(w-v)$ where h is strictly convex. Then there is a unique PM-compatible scheme S_y whose coupling attains the optimal cost $L(\mathbb{P},\mathbb{Q},c)$.

***Remark** 2:* It is known [10] that when $\mathsf{W} = [0,1]$: for **any** strictly convex $h(\cdot)$, the unique optimal map S_y is given by (13); this is Shayevitz & Feder's posterior matching scheme [2]. This property was recently used within the context of characterizing smooth properties of the optimal solution to Witsenhausen's counterexample [12].

In general, when $\dim(\mathsf{W}) > 1$, the aforementioned remark doesn't hold: as we shall see, different strictly convex cost functions induce optimal different maps.

A. The Brenier Map for Quadratic Cost

We now imagine $h(w,v) = |w-v|^2$. For this case, we have Brenier's theorem, which characterizes a second-order differential equation with a unique solution.

***Theorem** 4.4 (Brenier's Theorem):* Suppose $\mathsf{W} \subset \mathbb{R}^d$, $\mathbb{P}$ and $\mathbb{Q}$ induce probability density functions $p(w)$ and $q(w)$ respectively with respect to the Lebesgue measure and $c(w,v) = |u-v|^2$. Then there exists a unique convex $\varphi \in L^2$ such that

$$S(w) = \bigtriangledown\varphi(w) \tag{23}$$

and the coupling for S attains the optimal cost $L(\mathbb{P},\mathbb{Q},c)$. The second-order partial differential equation that arises from combining (23) and (22) is termed the Monge-Ampère Equation and can be solved using finite-element methods. See Theorem 2.1.5 of [15] for proof of the above theorem.

Note that the 1-dimensional case of the PM scheme, as developed in [2], is a special case of this framework [16, Example 3.2.14]. We also note that the connection between the PM scheme and OT is desirable for algorithmic purposes: there are well-known computational algorithms that can provide numerical solutions, as shown by previous studies in medical imaging [17], [18] and fluid dynamics [19], [20] etc.

See Figure 3 for an example of the Brenier map when $\mathsf{W} = [0,1]^2$. In this setting, $\mathsf{X} = \{1,2,3,4\}$ and $\phi(\cdot)$ maps $w \in [0,1]^2$ to one of the four quadrants it lies in. After observing the channel output $Y_1 = y$, the posterior distribution $f_{W_1|Y_1=y}$ is piece-wise constant over the four quadrants, and the values are given in Figure 3(L). The Brenier map S_y^* for which $S_y^* \# f_{W_1|Y_1=y} = f_{W_2|Y_1=y} = f_W$ is given in Figure 3(R).

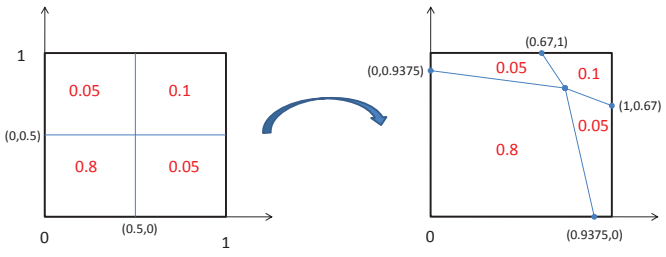

Fig. 3. (L): the posterior distribution $f_{W_1|Y_1=y}$. (R): the Brenier map S_y for which $f_{W_2|Y_1=y} \equiv f_W$.

B. The Knothe-Rosenblatt Map

We now consider an alternative cost function, of the form

$$c_\epsilon(w,v) = \sum_{k=1}^{d} \alpha_k(\epsilon)|w[k]-v[k]|^2 \tag{24}$$

where the coefficients $\alpha_k(\epsilon)$ are parametrized by ϵ. It can be shown [21] that if in the limit as $\epsilon \to 0$,

$$\frac{\alpha_k(\epsilon)}{\alpha_{k+1}(\epsilon)} \to 0, \tag{25}$$

then the Knothe-Rosenblatt coupling S_y, which is PM-compatible, attains $L(\mathbb{P},\mathbb{Q},c)$.

***Lemma** 4.5:* for $\mathsf{W} = [0,1]^d$, the Knothe-Rosenblatt (KR) map given by

$$w_{i+1}[1] = F_{W_i[1]|Y_i=y_i}(w_i[1]) \tag{26}$$
$$w_{i+1}[2] = F_{W_i[2]|Y_i=y_i,W_{i+1}[1]=w_{i+1}}(w_i[2]) \tag{27}$$
$$\dots \quad \dots$$

satisfies the Jacobian equation (11), elicits a PM-compatible scheme, and attains $L(\mathbb{P},\mathbb{Q},c)$ in (21) for the class of cost functions given in (24) for which (25) holds.

The proof is straightforward and can be found in [11]. Note that the Jacobian of the map is triangular. As such, preference is given more to certain axes of the message point as compared to the other. See Figure 4 for an example of the KR map when $\mathsf{W} = [0,1]^2$. We now provide a summary of the differences

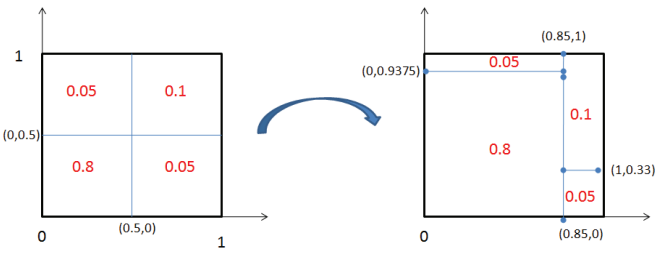

Fig. 4. (L): the posterior distribution $f_{W_1|Y_1=y}$. (R): the KR map S_y for which $f_{W_2|Y_1=y} \equiv f_W$.

between the Brenier scheme and the K-R scheme.

Brenier Transportation Map	KR Transportation Map
Cost function gives equal weight to all dimensions.	Cost function gives unequal weight to different dimensions
Need to solve a second-order PDE	Easy to implement – Closed form
Jacobian becomes Monge-Ampere PDE	Jacobian becomes upper-triangular matrix

V. Example: Multivariate Gaussian Case

In this section we will take two routes to solve for the optimal mapping in an example case of a multi-antenna additive Gaussian noise with causal feedback. We consider the case where the covariance matrices of the noise and the capacity achieving inputs are not necessarily the identity matrix. We will find optimal couplings with respect to the Brenier cost and the Knothe-Rosenblatt cost, and show that in both cases, the schemes achieve capacity - although in very different ways. The former scheme will be shown to be the multi-dimensional analogue to the 'innovations' scheme by Schalkwijk and Kailath[6], while the latter can be interpreted from an 'onion-peeling' perspective.

Suppose we have a multi-antenna communication problem with feedback. As such, we model $\mathsf{X} = \mathsf{Y} = \mathbb{R}^d$, where d is the number of transmit antennas and also the number of receive antennas. Naturally, to develop a PM-compatible scheme, it is desirable for $\Phi : \mathsf{W} \to \mathsf{X}$ to be an invertible map for which $\dim(\mathsf{W}) = \dim(\mathsf{X})$. As such, we model $\mathsf{W} = [0,1]^d$. We model the Gaussian noise as $Z_i \sim \mathscr{N}(0, \Sigma_N) \in \mathbb{R}^d$ where σ_N is not necessarily diagonal. The received signal is given by $Y_i = X_i + Z_i$.

See Figure 5.

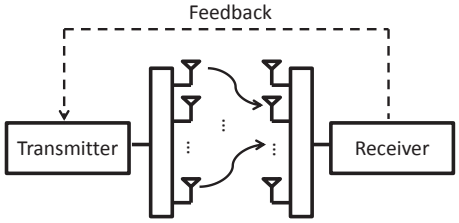

Fig. 5. Feedback communication over a Gaussian MIMO channel with feedback. $\dim(\mathsf{X}) > 1$.

Under an average power constraint, the optimal input distribution is given by $P_X = \mathscr{N}(0, \Sigma_X)$ for some σ_X. Note that $P_{X_1|Y_1} = \mathscr{N}(\mathbb{E}(X_1|Y_1), \Sigma_{X|Y})$. It can be directly shown [22, Prop 3.4.4] that $\mathbb{E}(X_1|Y_1) = \Sigma_X \Sigma_Y^{-1} Y_1$ and $\Sigma_{X|Y} = \Sigma_X - \Sigma_X \Sigma_Y^{-1} \Sigma_X$. From Corollary 3.2, it follows that it is our desire to construct a PM-compatible map $\bar{S}_y \equiv \phi \circ S_y \phi^{-1}$ such that $\bar{S}_y \# P_{X_1|Y_1} = P_X$.

A. The Optimal Brenier Map

The OT perspective dictates that we need to find a diffeomorphism $\bar{S}_y : \mathbb{R}^d \to \mathbb{R}^d$ that generates $X_{i+1} = \bar{S}_{Y_i}(X_i)$ which achieves

$$\inf_{\bar{S}_y} \int_{x \in \mathbb{R}^d} |x - S_y(x)|^2 P_{X_1|Y_1=y}(dx) \tag{28}$$

under the constraint that $\bar{S}_y \# P_{X_1|Y_1=y} = P_X$. Theorem 3.4.1 of [16] gives the solution of this OT problem (see [23] for original proof):

$$\begin{aligned} X_{i+1} &= \bar{S}_{Y_i}(X_i) = \alpha(X_i - \Sigma_X \Sigma_Y^{-1} Y_i) \qquad (29) \\ \alpha &= \Sigma_X^{\frac{1}{2}} \left(\Sigma_X^{\frac{1}{2}} \Sigma_{X|Y} \Sigma_X^{\frac{1}{2}} \right)^{-\frac{1}{2}} \Sigma_X^{\frac{1}{2}} \end{aligned}$$

Alternatively, using MMSE estimation theory and Corollary 3.2, we can alternatively find the Brenier map by first positing that

$$X_{i+1} = \alpha(X_i - \beta Y_i). \tag{30}$$

with unknown α and β. $\beta = \Sigma_X \Sigma_Y^{-1}$ ensures that $X_{i+1} \perp\!\!\!\perp Y_i$ (by joint Gaussianity and MMSE estimation), so we only need to find an invertible α. Because all variables are jointly Gaussian, and since both sides of (30) have zero-means, we can simply operate on covariance matrices.

$$\Sigma_X = \alpha(\Sigma_{X|Y})\alpha^T = \alpha(\Sigma_X - \Sigma_X \Sigma_Y^{-1} \Sigma_X)\alpha^T \tag{31}$$

It can be verified that this leads to the same linear algebra problem encountered in the OT formulation solved by Olkin and Pukelsheim [23]:

$$\alpha = \Sigma_X^{\frac{1}{2}} \left(\Sigma_X^{\frac{1}{2}} \Sigma_{X|Y} \Sigma_X^{\frac{1}{2}} \right)^{-\frac{1}{2}} \Sigma_X^{\frac{1}{2}}.$$

As such, it follows that the optimal Brenier map is the d-dimensional Schalkwijk-Kailath scheme [6]. This d-dimensional scheme was also used to prove fundamental limits of control over noisy channels in [8].

***Remark** 3:* It can be easily verified that the 1-dimensional case of [2, eqn. 22] that $X_{i+1} = \sqrt{1+\mathrm{SNR}}(X_i - \frac{\mathrm{SNR}}{1+\mathrm{SNR}} Y_i)$ is a natural derivation of the results (29) in higher dimension spaces.

B. The Optimal KR map

For the Knothe-Rosenblott cost, the optimal map in the $2-D$ Gaussian case is given by:

$$\begin{aligned} X_{i+1}[1] &= \beta_1 \left(X_i[1] - \mathbb{E}\left[X_i[1] | Y_i \right] \right) \qquad (32) \\ X_{i+1}[2] &= \beta_2 \left(X_i[2] - \mathbb{E}\left[X_i[2] | Y_i \right] \right) + \beta_3 X_{i+1}[1] \qquad (33) \end{aligned}$$

See [11] for the $\beta_1, \beta_2, \beta_3$ constants. This can be naturally extended for arbitrary $d > 1$. Note that the essence of this scheme is to first decode the first dimension of W, followed by the second given knowledge of the first, etc. Here, we are expanding $I(W;Y) = I(W[1];Y) + I(W[2];Y|W[1])$. Thus, this has the potential to be applicable to unequal error protection scenarios where the first dimension of W has more important information than the second.

Conclusion

We have demonstrated via the theory of optimal transportation that the posterior matching scheme can be generalized to arbitrary dimensions. In short, the cost function in the theory of the optimal transportation determines which PM-compatible map to select. Such a map can be implemented in a dynamical system encoder with low complexity - and is guaranteed to maximize mutual information for any n. Different cost functions can be applicable for different purposes. For example, the Knothe-Rosenblatt cost - and its induced map - has the potential to be applicable for unequal error protection scenarios. Lastly, we would like to comment that multi-dimensional messages can pertain to different messages at encoders or decoders for multi-terminal information theory

problems. For example, we can treat the two independent messages on $[0,1]$ in the physically degraded broadcast with feedback as a single message in $[0,1]^2$ and recover a Knothe-Rosenblatt type of posterior matching scheme [24, eqn 4].

The authors would like to thank Nihar Jindal for originally suggesting a generalization of posterior matching to higher dimensions to address multi-antenna Gaussian channels. The authors initially pondered exploring a relationship between optimal transportation and posterior matching after attending a presentation of Yihong Wu presentation pertaining to [12] in February 2011. We would also like to thank Siva Gorantla for useful discussions.

REFERENCES

[1] R. McEliece, *The theory of information and coding*. Cambridge Univ Press, 2002.

[2] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1186–1222, 2011.

[3] S. K. Gorantla and T. P. Coleman, "Information-Theoretic Viewpoints on Optimal Causal Coding-Decoding Problems," *IEEE Transactions on Information Theory*, submitted Jan 2011.

[4] ——, "Equivalence between reliable feedback communication and nonlinear filter stability," in *IEEE International Symposium on Information Theory*, St. Petersburg, Russia, August 2011.

[5] T. M. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[6] J. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback–I: No bandwidth constraint," *Information Theory, IEEE Transactions on*, vol. 12, no. 2, pp. 172–182, 1966.

[7] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Transactions on Information Theory*, vol. 9, no. 3, pp. 136–143, 1963.

[8] N. Elia, "When Bode meets Shannon: Control-oriented feedback communication schemes," *IEEE Transactions on Automatic Control*, vol. 49, no. 5, pp. 1477–1488, 2004.

[9] C. Omar, A. Akce, M. Johnson, T. Bretl, R. Ma, E. Maclin, M. McCormick, and T. P. Coleman, "A Feedback Information-Theoretic Approach to the Design of Brain-Computer Interfaces," *Int'l Journal on Human-Computer Interaction*, January 2011, special issue on Current Trends in Brain-Computer Interface (BCI) Research.

[10] C. Villani, *Optimal transport: old and new*. Springer Verlag, 2009.

[11] M. Rosenblatt, "Remarks on a multivariate transformation," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 470–472, 1952.

[12] Y. Wu and S. Verdú, "Witsenhausen's counterexample: a view from optimal transport theory," in *IEEE Conference on Decision and Control (CDC)*, December 2011, to appear.

[13] G. Monge, *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

[14] L. Kantorovich, "On mass transportation," in *Dokl. Akad. Nauk. SSSR*, vol. 37, 1942, pp. 227–229.

[15] C. Villani, *Topics in optimal transportation*. American Mathematical Society, 2003.

[16] S. Rachev and L. Rüschendorf, *Mass transportation problems*. Springer Verlag, 1998, vol. 1.

[17] S. Haker, A. Tannenbaum, and R. Kikinis, "Mass preserving mappings and image registration," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*. Springer, 2010, pp. 120–127.

[18] E. Haber, T. Rehman, and A. Tannenbaum, "An efficient numerical method for the solution of the L2 optimal mass transfer problem," *SIAM journal on scientific computing: a publication of the Society for Industrial and Applied Mathematics*, vol. 32, no. 1, p. 197, 2010.

[19] J. Benamou and Y. Brenier, "A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem," *Numerische Mathematik*, vol. 84, no. 3, pp. 375–393, 2000.

[20] G. Liao, X. Cai, J. Liu, X. Luo, J. Wang, and J. Xue, "Construction of differentiable transformations," *Applied Mathematics Letters*, vol. 22, no. 10, pp. 1543–1548, 2009.

[21] F. Santambrogio, A. Galichon, and G. Carlier, "From Knothe's transport to Brenier's map and a continuation method for optimal transport," *SIAM Journal on Mathematical Analysis*, vol. 41, no. 6, p. 2554, 2010.

[22] B. Hajek, *An Exploration of Random Processes for Engineers*. http://www.ifp.illinois.edu/ hajek/Papers/randomprocJan09.pdf, 2009.

[23] F. Olkin *et al.*, "The distance between two random vectors with given dispersion matrices," *Linear Algebra and its Applications*, vol. 48, pp. 257–263, 1982.

[24] S. Gorantla and T. Coleman, "A stochastic control approach to coding with feedback over degraded broadcast channels," in *IEEE Conference on Decision and Control*, 2010.