Witsenhausen's Counterexample: a View from Optimal Transport Theory

Yihong Wu and Sergio Verdú

***Abstract*— We formulate Witsenhausen's counterexample in stochastic control as an optimization problem involving the quadratic Wasserstein distance and the minimum mean-square error. Classical results are recovered as immediate consequences of transport-theoretic properties. New results and bounds on the optimal cost are also obtained. In particular, we show that the optimal controller is a strictly increasing function with a real analytic left inverse.**

I. Introduction

In [1] Witsenhausen constructed a linear quadratic Gaussian (LQG) team problem with non-classical information structure and showed that the linear controller is not necessarily optimal. This serves as a counterexample to the conjectured optimality of linear controllers in LQG problems.

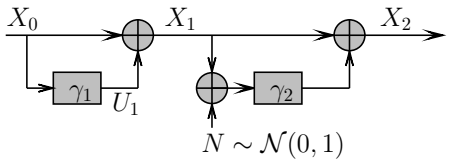

Fig. 1. Witsenhausen's decentralized stochastic control problem.

As illustrated in Fig. 1, Witsenhausen's counterexample is a two-stage decentralized stochastic control problem, where the goal is to minimize the weighted average control cost $k^2\mathbb{E}\left[U_1^2\right] + \mathbb{E}\left[X_2^2\right]$ over all pairs of controllers γ_1 and γ_2 that are Borel measurable. According to the notation in [1], let $f(x) = \gamma_1(x) + x$ and $g(x) = \gamma_2(x)$ and denote the weighted control cost achieved by (f, g) by

$$J(f,g) = k^2\mathbb{E}\left[(f(X_0) - X_0)^2\right] \tag{1}$$
$$+ \mathbb{E}\left[(f(X_0) - g(f(X_0) + N))^2\right], \tag{2}$$

where $N \sim \mathcal{N}(0,1)$ is independent of X_0, whose distribution is fixed and arbitrary.

For a given f, the optimal g is the minimum mean-square error (MMSE) estimator of $f(X_0)$, i.e., the conditional mean, given the noisy observation:

$$g_f^*(\cdot) = \mathbb{E}\left[f(X_0)|f(X_0) + N = \cdot\right]. \tag{3}$$

Therefore

$$\min_g J(f,g) = k^2\mathbb{E}\left[(f(X_0) - X_0)^2\right] + \mathsf{mmse}(f(X_0), 1). \tag{4}$$

Yihong Wu and Sergio Verdú are with the Department of Electrical Engineering, Princeton University, Princeton NJ 08540, USA, yihongwu@princeton.edu, verdu@princeton.edu

where

$$\mathsf{mmse}(X, \sigma^2) \triangleq \min_g \mathbb{E}\left[(X - g(\sigma X + N))^2\right] \tag{5}$$
$$= \mathbb{E}\left[\mathsf{var}(X|\sigma X + N)\right]. \tag{6}$$

Since (6) only depends on the distribution of X, we also denote $\mathsf{mmse}(P_X, \sigma^2) = \mathsf{mmse}(X, \sigma^2)$. The properties of the MMSE functional as a function of the input distribution and the signal-to-noise-ratio have been studied in [2] and [3] respectively.

Next we define the optimal cost functional. Introduce a scale parameter by letting $X_0 = \sigma X$ with X distributed according to some probability measure P. Denote the optimal cost by

$$J^*(k^2, \sigma^2, P)$$
$$\triangleq \inf_{f,g} J(f,g) \tag{7}$$
$$= \inf_f k^2\mathbb{E}\left[(X_0 - f(X_0))^2\right] + \mathsf{mmse}(f(X_0), 1) \tag{8}$$
$$= \inf_f k^2\sigma^2\mathbb{E}\left[(X - f(X))^2\right] + \sigma^2\mathsf{mmse}(f(X), \sigma^2). \tag{9}$$

The optimal affine cost is denoted by $J_{\mathrm{a}}^*(k^2, \sigma^2, P)$, defined as the infimum in (7) with f and g restricted to affine functions. Direct computation shows that (see [1, p. 141])

$$J_{\mathrm{a}}^*(k^2, \sigma^2, P) = \min_{\lambda \geq 0} k^2\sigma^2(1-\lambda)^2\,\mathsf{var}P + \frac{\lambda^2\sigma^2\,\mathsf{var}P}{1 + \lambda^2\sigma^2\,\mathsf{var}P}. \tag{10}$$

When the input is standard Gaussian, we simplify

$$J^*(k^2, \sigma^2) \triangleq J^*(k^2, \sigma^2, \mathcal{N}(0,1)). \tag{11}$$

The same convention also applies to $J_{\mathrm{a}}^*(k^2, \sigma^2)$.

The above is the usual formulation of the Witsenhausen's problem. In [1], it is shown that optimal controller that attains the infimum in (9) exists for arbitrary input distribution and is a non-decreasing function. Moreover, for Gaussian input distribution, Witsenhausen showed that

$$J^*(k^2, \sigma^2) < J_{\mathrm{a}}^*(k^2, \sigma^2), \tag{12}$$

holds in the regime of $k = \frac{1}{\sigma}$ and sufficiently large σ. The proof involves showing that a two-point quantizer $f(x) = \mathrm{sgn}(x)$ yields strictly smaller cost than the best affine controller. This observation has been further extended: [4] lowered the cost by letting $f(x) = \sqrt{\frac{2}{\pi}}\mathrm{sgn}(x)$, while [5] showed that the ratio between the optimal cost and the optimal affine cost can be made unbounded by using successively finer quantizers of the Gaussian distribution, i.e.,

$$\lim_{\sigma \to \infty} \frac{J_{\mathrm{a}}^*(\sigma^{-2}, \sigma^2)}{J^*(\sigma^{-2}, \sigma^2)} = \infty. \tag{13}$$

Numerical algorithms that provide upper bounds on the optimal cost have been proposed using neural networks [6], hierarchical search [7], learning approach [8], etc. Based on information-theoretical ideas, [9], [10] developed upper and lower bounds that are within a constant factor using lattice quantization and joint-source-channel coding converse respectively. For a comprehensive review see [11], [12]. Determining the optimal controller remains an open problem.

In this paper, we take a new optimal transport theoretic approach and give a concise formulation of Witsenhausen's counterexample in terms of the *quadratic Wasserstein distance* and the MMSE functional. Capitalizing on properties of the optimal transport mapping, Witsenhausen's classical results can be recovered and extended with much simpler proofs. Moreover, we show that

1) For Gaussian input, the optimal controller is a strictly increasing function with a *real analytic* left inverse. Based on the numerical evidence in [6], it was believed that piecewise affine controller is optimal (see for example [7, p. 384] and the conjecture in [10]). However, our result shows that this is not the case.
2) For Gaussian input, (12) holds for any $k < 0.564$ and sufficiently large σ. This improves the result in [1] which only applies to the regime of $k = \frac{1}{\sigma}$.
3) For any input distribution, the best affine controller is asymptotically optimal in the weak-signal regime ($\sigma \to 0$).

Various properties and bounds on the optimal cost are also obtained.

II. Optimal transport theory

Optimal transport theory deals with the most economic way of distributing supply to meet the demand. Consider the following illustrative example [13, Chapter 3]: two bakeries are located at x_1 and x_2, producing three and four units of bread each day respectively. Three cafés, located at y_1, y_2 and y_3, consume two, four and one units of bread daily respectively. Assuming the transport cost is proportional to the distance and the amount of bread, the question is how to transport the bread from bakeries to cafés so as to minimize the total cost. One feasible transport plan is illustrated in Fig. 2, whose total cost is $2|x_1-y_1|+|x_1-y_3|+4|x_2-y_2|$.

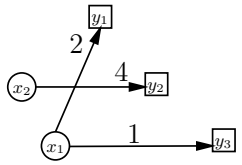

Fig. 2. Example of a transport plan.

Monge-Kantorovich's probabilistic formulation of the optimal transportation problem is as follows: Given probability measures P and Q and a cost function $c : \mathbb{R}^2 \to \mathbb{R}$, define

$$\inf_{P_{XY}} \{\mathbb{E}\left[c(X,Y)\right] : P_X = P, P_Y = Q\} \tag{14}$$

where the infimum is over all joint distributions (couplings) of (X, Y) with prescribed marginals P and Q. To see the relationship between (14) and the optimal transport problem, note that the example in Fig. 2 corresponds to

$$P_X = \frac{3}{7}\delta_{x_1} + \frac{4}{7}\delta_{x_2},\ P_Y = \frac{2}{7}\delta_{y_1} + \frac{4}{7}\delta_{y_2} + \frac{1}{7}\delta_{y_3} \tag{15}$$

$$P_{Y|X=x_1} = \frac{2}{3}\delta_{y_1} + \frac{1}{3}\delta_{y_3},\ P_{Y|X=x_2} = \delta_{y_2}, \tag{16}$$

where δ_x is the Dirac measure (point mass) at x. The transportation cost normalized by the total amount of bread is exactly $\mathbb{E}\left[c(X,Y)\right]$ with $c(x,y) = |x-y|$.

With $c(x,y) = (x-y)^2$, the quadratic Wasserstein distance [13, Chapter 6] is defined as follows:

Definition 1. The *quadratic Wasserstein space* on $\mathbb{R}$ is defined as the collection of all Borel probability measures with finite second moments, denoted by $\mathcal{P}_2(\mathbb{R})$. The *quadratic Wasserstein distance* is a metric on $\mathcal{P}_2(\mathbb{R})$, defined for $P, Q \in \mathcal{P}_2(\mathbb{R})$ as

$$W_2(P,Q) = \inf_{P_{XY}} \{\|X-Y\|_2 : P_X = P, P_Y = Q\}, \tag{17}$$

where $\|X-Y\|_2 \triangleq \sqrt{\mathbb{E}\left[(X-Y)^2\right]}$.

The W_2 distance metrizes convergence in distribution and of second-order moments, i.e., $W_2(P_{X_k}, P_X) \to 0$ if and only if $X_k \xrightarrow{\mathrm{D}} X$ and $\mathbb{E}[X_k^2] \to \mathbb{E}[X^2]$.

Let F_P and F_P^{-1} denote the cumulative distribution function (CDF) and quantile function (functional inverse of the CDF [14, Exercise II.1.18]) of P respectively. The infimum in (17) is attained by a unique coupling P^*_{XY}, which can be represented by $X = \mathsf{F}_P^{-1}(U)$ and $Y = \mathsf{F}_Q^{-1}(U)$, for some U uniformly distributed on [0,1]. The distribution function of P^*_{XY} is given by $F^*(x,y) = \min\{\mathsf{F}_P(x), \mathsf{F}_Q(y)\}$ [15, Section 3.1]. Therefore, the W_2 distance is simply the L_2 distance between the respective quantiles [16]:

$$W_2(P,Q) = \left\|\mathsf{F}_P^{-1} - \mathsf{F}_Q^{-1}\right\|_2. \tag{18}$$

If P is atomless, the optimal coupling $P^*_{Y|X}$ is *deterministic*, i.e., $Y = f(X)$ with

$$f = \mathsf{F}_Q^{-1} \circ \mathsf{F}_P. \tag{19}$$

The following properties of the Wasserstein distance are relevant to our subsequent analysis:

Lemma 1.

(a) $(P,Q) \mapsto W_2(P,Q)$ *is weakly lower semi-continuous.*
(b) For any fixed P, $Q \mapsto W_2^2(P,Q)$ *is convex.*
(c) $W_2(P_{aX}, P_{aY}) = |a| W_2(P_X, P_Y)$.
(d) $W_2^2(P_{X+x}, P_{Y+y}) =$
$W_2^2(P_X, P_Y) + (x-y)^2 + 2(\mathbb{E}\left[X\right] - \mathbb{E}\left[Y\right])(x-y)$.
(e)

$$\left|\sqrt{\mathsf{var}X} - \sqrt{\mathsf{var}Y}\right|^2 \le W_2^2(P_X, P_Y) - (\mathbb{E}\left[X\right] - \mathbb{E}\left[Y\right])^2 \tag{20}$$

$$\le \mathsf{var}X + \mathsf{var}Y. \tag{21}$$

(f) For any strictly increasing function $f : \mathbb{R} \to \mathbb{R}$, $W_2(P_X, P_{f(X)}) = \|X - f(X)\|_2$. In particular, for all $a > 0$, $W_2(P_X, P_{aX}) = |a-1| \|X\|_2$.

(g) Let $m_i(P)$ denote the i^{th} moment of P. Then

$$\min_{Q:\, \mathsf{var} Q \le \sigma^2} W_2(P, Q) = \left|\sqrt{\mathsf{var} P} - \sigma\right|, \tag{22}$$

attained by the affine coupling $f(x) = m_1(P) + \frac{\sigma}{\sqrt{\mathsf{var} X}}(x - m_1(P))$.

(h) $W_2(P * Q, P' * Q) \le W_2(P, P')$.

(i) $W_2^2(P_X, \delta_x) = \mathsf{var} X + (\mathbb{E}[X] - x)^2$.

Proof. (a): [17, Proposition 7.1.3].

(b): Let λ_1 and λ_2 be the optimal coupling of (P, Q_1) and (P, Q_2) respectively. For any $0 < \alpha < 1$, $\lambda = \alpha\lambda_1 + (1-\alpha)\lambda_2$ is a coupling for $(P, \alpha Q_1 + (1-\alpha)Q_2)$. Therefore

$$\begin{aligned} & W_2^2(P, \alpha Q_1 + (1-\alpha)Q_2) \\ \le & \int (x-y)^2 \lambda(\mathrm{d}x, \mathrm{d}y) \qquad (23) \\ = & \alpha W_2^2(P, Q_1) + (1-\alpha) W_2^2(P, Q_2). \qquad (24) \end{aligned}$$

(c): $\|aX - aY\|_2 = |a| \|X - Y\|_2$.

(d): Direct calculation.

(e): The lower bound is due to the triangle inequality and the upper bound is given by an independent coupling.

(f): By (19).

(g): By (d), (e) and (f).

(h): For any coupling $P_{XX'}$ of (P, P'), $P_{X+Z, X'+Z}$ is a coupling $P_{XX'}$ of $(P * Q, P' * Q)$, where $Z \sim Q$ is independent of (X, X').

(i): Direct calculation. □

In view of Lemma 1(d) and (f), the W_2 distance between Gaussian distributions is given by

$$W_2^2(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2, \tag{25}$$

attained by an affine coupling: $Y = \mu_2 + \frac{\sigma_2}{\sigma_1}(X - \mu_1)$.

III. Transport-theoretic formulation of Witsenhausen's counterexample

We reformulate Witsenhausen's counterexample in terms of the Wasserstein distance by allowing randomized controllers, i.e., relaxing the controller from a deterministic function f to a random transformation (transition probability kernel) $P_{Y|X}$.[1] In fact, a concavity argument shows that such relaxation incurs no loss of generality. Indeed, for a fixed g, the weighted cost $J(P_{Y|X}, g)$ is *affine* in $P_{Y|X}$. Therefore the pointwise infimum $\inf_g J(P_{Y|X}, g)$ is *concave* in $P_{Y|X}$, whose minimum occurs on extremal points, i.e., deterministic controllers. The fact that randomized policy does *not* help is standard in stochastic decision problems (e.g., [18, Section 8.5] or [19, Theorem 4.1]).

[1]This is in the same spirit as Kantorovich's generalization of Monge's original optimal transport problem, which allows only deterministic couplings in (14).

Based on the above reasoning and (9), we obtain a new formulation of Witsenhausen's problem as:

$$\begin{aligned} J^*(k^2, \sigma^2, P) &= \sigma^2 \inf_{P_{Y|X}} k^2 \mathbb{E}\left[(X-Y)^2\right] + \mathsf{mmse}(Y, \sigma^2) \qquad (26) \\ &= \sigma^2 \inf_{Q} \left\{k^2 W_2^2(P, Q) + \mathsf{mmse}(Q, \sigma^2)\right\}, \qquad (27) \end{aligned}$$

which involves minimizing the MMSE penalized by the W_2 distance.

Related problems to (27) have been studied in the partial differential equations community. For example, maximizing the differential entropy is considered in [20], [21]:

$$\inf_Q \left\{k^2 W_2^2(P, Q) - h(Q)\right\}, \tag{28}$$

where $h(Q) = -\int \log q \, \mathrm{d}Q$ denotes the differential entropy of probability measure Q with density q. Solving (28) gives a variational scheme to compute discretized approximation to the solution of the Fokker-Planck equation [20]. Note that for Gaussian P, the infimum in (28) is attained by a Gaussian Q [21, p. 821]. This is because for a given variance, a Gaussian Q minimizes $W_2^2(P, Q)$ and maximizes $h(Q)$ simultaneously. Another problem involving energy minimization is studied in [22]:

$$\inf_Q \left\{k^2 W_2^2(P, Q) + \int \Psi \, \mathrm{d}Q\right\}. \tag{29}$$

Note that (28) and (29) are both convex optimization problems, because $-h(Q)$ and $\int \Psi \, \mathrm{d}Q$ are *convex* and *affine* in Q respectively. Comparing (28) and (29) with (27), we see that the difficulty in Witsenhausen's problem lies in the *concavity* of $Q \mapsto \mathsf{mmse}(Q, \sigma^2)$ [2, Theorem 2], which results in the non-convexity of the optimization problem.

IV. Optimal controller

A. Existence

We give a simple proof of the existence of optimal controller:

Theorem 1. *For any P, the infimum in (27) is attained.*

Proof. In view of Lemma 1(g), Q can be restricted to the *weakly compact* subset $\{Q : m_2(Q) \le 4 m_2(P)\}$ of $\mathcal{P}_2(\mathbb{R})$, where $m_2(\cdot)$ denotes the second-order moment. By Lemma 1(a), $Q \mapsto W_2(P, Q)$ is *weakly lower semicontinuous*, while $Q \mapsto \mathsf{mmse}(Q, \sigma^2)$ is *weakly continuous* for any $\sigma > 0$ [2, Theorem 7]. The existence of the minimizer of (27) then follows from the fact that lower semicontinuous functions attain infimum on compact set. □

The above proof is much simpler than Witsenhausen's original argument [1, Theorem 1], which involves proving that an infimizing sequence of controller converges pointwise and the limit is optimal. Note that Theorem 1 also holds for non-Gaussian noise, as long as the noise has a continuous and bounded density which guarantees the weak continuity of MMSE [2, Theorem 7].

Output distribution Q	Controller f
Gaussian	affine
discrete	piecewise constant
atomless	strictly increasing
bounded supported	bounded
symmetric	odd
has smooth density	smooth

TABLE I

RELATIONSHIP BETWEEN OUTPUT DISTRIBUTION AND CONTROLLER.

B. Structure of the optimal controller

Any optimal controller is an optimal transport mapping from P to the optimal Q. In view of (19), the optimal controller is an *increasing* function. In case of $P = \mathcal{N}(0,1)$, the optimal controller is given by

$$f = \mathsf{F}_Q^{-1} \circ \Phi, \tag{30}$$

where Φ denotes the standard Gaussian CDF. As summarized in Table I, various properties of the controller f can be *equivalently* recast as constraints on the output distribution Q. For example, using only affine controllers is equivalent to restricting Q to Gaussian distributions. Observe that for Gaussian P, there is an incentive for using non-linear control (equivalently non-Gaussian Q). By Lemma 1(g), among all distributions with the same variance, Gaussian Q *minimizes* the W_2 distance to P but *maximizes* the MMSE [3, Proposition 15]. Therefore it is possible the optimal Q is non-Gaussian.

C. Regularity of optimal controller

It is known that the optimal g as a MMSE estimator is real analytic [1, Lemma 3]. The following result shows that the optimal f is a strictly increasing piecewise real analytic function with a real analytic left inverse. According to the identity theorem of real analytic functions [23, Theorem 9.4.3, p. 208], piecewise affine functions do not have analytic left inverses. Therefore we conclude that *piecewise constant* or *piecewise affine* controllers cannot be optimal, disproving a conjecture in [10, p. 21]. Nevertheless, since MMSE is weakly continuous, the optimal cost can be approached arbitrarily close by restricting Q to any weakly dense subset of $\mathcal{P}_2(\mathbb{R})$ (e.g., discrete distributions, Gaussian mixtures, etc.) or restricting the controller f to any dense family of $L^2(\mathbb{R}, P)$ (e.g., piecewise constant or affine functions).

Theorem 2. *Let P has a real analytic strictly positive density. Then*

- *Any optimal Q for (27) has a real analytic density and unbounded support, with the same mean as P and variance not exceeding $\mathsf{var}P + \frac{4}{k^2\sigma^2}$.*
- *Any optimal controller f is a strictly increasing unbounded piecewise real analytic function with a real analytic inverse.*

Proof. For notational convenience, assume that $\sigma = 1$. Let Q be an minimizer of (27) and $Y = f(X)$ is the associated optimal coupling. Proceeding as in the proof of [20, Theorem 5.1], fix $\tau \in \mathbb{R}$ and $\xi \in C_c^\infty(\mathbb{R})$ arbitrarily. Perturb Y along the direction of ξ by letting

$$Y_\tau = f(X) + \tau\,\xi(X) \tag{31}$$

and $Q_\tau = P_{Y_\tau}$. Then

$$\begin{aligned} &W_2^2(P, Q_\tau) - W_2^2(P, Q) \\ &\le \mathbb{E}\left[(X - (f+\tau\xi)(X))^2\right] - \mathbb{E}\left[(X - f(X))^2\right] \qquad (32)\\ &= 2\tau\,\mathbb{E}\left[\xi(X)(f(X) - X)\right] + \tau^2 \mathbb{E}\left[\xi^2(X)\right]. \qquad (33)\end{aligned}$$

It can be shown that the first-order variation on the MMSE is

$$\begin{aligned} &\mathsf{mmse}(Q_\tau, \sigma^2) - \mathsf{mmse}(Q, \sigma^2) \\ &= -\tau\,\mathbb{E}\left[(\varphi' * (\eta^2 + 2\eta')) \circ f(X)\xi(X)\right] + o(\tau). \qquad (34)\end{aligned}$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}$ denotes the standard normal density, $\eta = \frac{g'}{g}$ is the *score* function of $Z = Y + N$ and

$$g(z) = \mathbb{E}\left[\varphi(z - N)\right] \tag{35}$$

is the density of Z. By the optimality of Q, we have

$$\begin{aligned} &2k^2\,\mathbb{E}\left[(f(X) - X)\xi(X)\right] \\ &\le \liminf_{\tau \downarrow 0} \frac{1}{\tau} k^2 (W_2^2(P, Q_\tau) - W_2^2(P, Q)) \qquad (36)\\ &\le \mathbb{E}\left[(\varphi' * (\eta^2 + 2\eta')) \circ f(X)\xi(X)\right], \qquad (37)\end{aligned}$$

where (36) and (37) follows from (33) and (34) respectively. Replacing τ by $-\tau$ in (36) and by the arbitrariness of ξ, the following variational equation holds P-a.e. (or equivalently Lebesgue-a.e.):[2]

$$2k^2(f - \mathsf{id}) = (\varphi' * (\eta^2 + 2\eta')) \circ f, \tag{38}$$

where $\mathsf{id}(x) = x$. In view of (19), f is right-continuous, which implies that (38) actually holds *everywhere*.

An immediate consequence of the variational equation is the regularity of the optimal controller. Let

$$h = \mathsf{id} - \frac{1}{2k^2}(\varphi' * (\eta^2 + 2\eta')). \tag{39}$$

Then

$$h \circ f = \mathsf{id}, \tag{40}$$

i.e., h is a left inverse of f. Therefore f is injective [24, Theorem I.1, p. 7], hence strictly increasing. Due to the analyticity of the Gaussian density, $\varphi' * (\eta^2 + 2\eta')$ is real analytic regardless of η [1, Lemma 2]. Thus h is also real analytic. Note that f has at most countably many discontinuities. We conclude that f is piecewise real analytic.[3] In view of the continuity of h, (40) implies that the range of f is unbounded.

[2]Directly perturbing the distribution of Y results in the same variational equation.

[3]Note that (40) alone does not imply that f is analytic. For a counterexample, consider the analytic function $g(x) = x^3 - x$. Let f be the inverse of g restricted on $|x| \ge 1$. Then (40) is satisfied but f has a discontinuity at 0. To prove the analyticity of f is equivalent to show that Q is supported on the entire real line.

Next we show that Q is absolutely continuous with respect to the Lebesgue measure. In view of Table I, the strict monotonicity of f implies that Q has no atom. Let $f^{-1}: f(\mathbb{R}) \to \mathbb{R}$ denote the inverse of f. Since $f^{-1} = \mathsf{F}_P^{-1} \circ \mathsf{F}_Q$, (40) implies that $\mathsf{F}_Q = \mathsf{F}_P \circ h$ holds on the entire range of f, whose closure is the support of Q. By assumption, F_P is a real analytic function. It follows that F_Q is also real analytic, i.e., Q has a density that is real analytic in the interior of its support.

To conclude the proof, we show an upper bound on $\mathsf{var}Q$. From (38), we have

$$X = Y - \frac{1}{2k^2}\varphi' * (\eta^2 + 2\eta') \circ Y, \text{ a.s.} \tag{41}$$

Without loss of generality, we assume that $\mathbb{E}[X] = 0$. Then $\mathbb{E}[Y] = 0$ in view of Lemma 1(d). Hence

$$\begin{aligned} & k^2(\mathsf{var}Y - \mathsf{var}X) \\ \le\ & \mathbb{E}\left[Y(\varphi' * (\eta^2 + 2\eta')) \circ Y\right] && (42) \\ =\ & \int \mathbb{E}\left[Y\varphi'(z - Y)\right](\eta^2 + 2\eta')(z)\mathrm{d}z && (43) \\ =\ & \int (g(z) + zg'(z) + g''(z))(\eta^2 + 2\eta')(z)\mathrm{d}z, && (44) \end{aligned}$$

Recall that $\eta = \frac{g'}{g}$ is the score of Z, and the Fisher information of Z is given by

$$J(Z) = \int g\eta^2 = -\int g\eta'. \tag{45}$$

Hence

$$\int g(\eta^2 + 2\eta')\mathrm{d}z = -J(Z) \tag{46}$$

Integrating by parts, we have

$$\begin{aligned} & \int zg'(z)(\eta^2 + 2\eta')(z)\mathrm{d}z \\ =\ & \int zg'(z)\eta^2(z)\mathrm{d}z + 2\int zg(z)\eta(z)\eta'(z)\mathrm{d}z && (47) \\ =\ & \int zg'(z)\eta^2(z)\mathrm{d}z - \int (zg(z))'\eta^2(z)\mathrm{d}z && (48) \\ =\ & -\int g(z)\eta^2(z)\mathrm{d}z && (49) \\ =\ & -J(Z), && (50) \end{aligned}$$

where we have used $\eta = \frac{g'}{g}$. Similarly,

$$\begin{aligned} & \int g''(\eta^2 + 2\eta')\mathrm{d}z \\ =\ & -\int \eta^2(g'' + g)\mathrm{d}z + \int \eta^2 g\mathrm{d}z + 2\int \frac{g''^2}{g}\mathrm{d}z && (51) \\ =\ & -\mathbb{E}\left[\eta^2(Y)\mathbb{E}\left[N^2|Z\right]\right] + J(Z) + 2\mathbb{E}\left[(\mathbb{E}\left[N^2|Z\right] - 1)^2\right] && (52) \\ \le\ & J(Z) + 2(\mathbb{E}\left[N^4\right] - 1) && (53) \\ =\ & J(Z) + 4. && (54) \end{aligned}$$

where (52) is due to (35), (45) and $\frac{g''}{g}(z) = \mathbb{E}\left[N^2|Z = z\right] - 1$, while (53) follows from Jensen's inequality. Combining (44), (46), (50) and (54), we have

$$\mathsf{var}Q \le \mathsf{var}P + \frac{4 - J(Z)}{k^2} \le \mathsf{var}P + \frac{4}{k^2}. \tag{55}$$

□

Remark 1. Combining the Crámer-Rao bound $J(Z) \ge \frac{1}{\mathsf{var}Z} = \frac{1}{1+\mathsf{var}Q}$ with the first inequality in (55) yields a better upper bound: $\mathsf{var}Q \le \frac{g(k^2, \sigma^2\mathsf{var}P)}{\sigma^2}$, where

$$g(u, v) = \frac{v + \frac{4}{u} - 1}{2} + \frac{1}{2}\sqrt{\left(v + \frac{4}{u} + 1\right)^2 - \frac{1}{u}}. \tag{56}$$

Remark 2. The variational equation (38) has been formally derived in [1, p. 140], where it is remarked that "this condition is of little use". However, combined with the structure of optimal controller as optimal transport map, interesting results can be deduced.

For Gaussian input, solutions to (38) always exist, namely optimal linear controllers. This has also been observed by Witsenhausen [1, Lemma 14]. In view of the analyticity result in Theorem 3, finding series approximations to the solution of (38) is a reasonable attempt to find good controllers. However, it can be shown that the only polynomial solution to (38) is affine.

V. Optimal cost

A. Properties

Theorem 3. *$P \mapsto J^*(k^2, \sigma^2, P)$ is concave, weakly upper semi-continuous and translation-invariant. Moreover,*

$$0 \le J^*(k^2, \sigma^2, P) \le \min\{k^2\sigma^2\mathsf{var}P, \sigma^2\,\mathsf{mmse}(P, \sigma^2)\} \le 1. \tag{57}$$

Proof. By (7), $J^*(k^2, \sigma^2, \cdot)$ is the pointwise infimum of affine functionals, hence concave. Weak semicontinuity follows from pointwise infimum of weak continuous functionals (see the proof of [2, Theorem 6]). The middle inequality in (57) follows from choosing Q to be either $\delta_{m_1(P)}$ or P. □

The following result gives a lower bound on the optimal cost of any symmetric distribution via the optimal cost of the Rademacher distribution (random sign) $B = \frac{1}{2}(\delta_1 + \delta_{-1})$, which has been explicitly determined in [1, Sec. 5] (see Fig. 3):

$$J^*(k^2, \sigma^2, B) = \min_{b \ge 0}\left\{k^2(b - \sigma)^2 + b^2\,\mathsf{mmse}(B, b^2)\right\} \tag{58}$$

where $b^2\,\mathsf{mmse}(B, b^2) = \sqrt{2\pi}a^2\varphi(a)\int \frac{\varphi(y)}{\cosh(ay)}\mathrm{d}y$.

Theorem 4. *For any symmetric P,*

$$\begin{aligned} & J^*(k^2, \sigma^2, P) \\ \ge\ & \sup_{\substack{Q, Q': \\ \frac{1}{2}(Q+Q') = P}} \sup_{\substack{P_{YY'}: \\ P_Y = Q, P_{Y'} = Q'}} \mathbb{E}\left[J^*(k^2, \sigma^2|Y - Y'|^2/4, B)\right] && (59) \\ \ge\ & \sup_{\substack{P_{YY'}: \\ P_Y = P_{Y'} = P}} \mathbb{E}\left[J^*(k^2, \sigma^2|Y - Y'|^2/4, B)\right]. && (60) \end{aligned}$$

The proof of Theorem 4 follows from writing a symmetric distribution as a *scale mixture* of the Rademacher distribution and concavity of the optimal cost. For symmetric P, choosing the coupling $Y' = -Y$ in (60) gives the the lower bound in [1, Theorem 3].

B. Monotonicity in signal power

Consider the following question: for a given input distribution P, does higher power necessarily require higher control cost, i.e., for fixed k^2 and P, is $J^*(k^2, \sigma^2, P)$ increasing in σ^2? Intuitively this should be true. However, any discrete input with finite variance serves as an counterexample (see Fig. 3 for binary input). To see this, by (57), $J^*(k^2, \sigma^2, P) \le \sigma^2 \,\mathsf{mmse}(P, \sigma^2)$, which vanishes as $\sigma \to 0$ or ∞.[4] Therefore $J^*(k^2, \cdot, P)$ cannot be monotone for any discrete P.

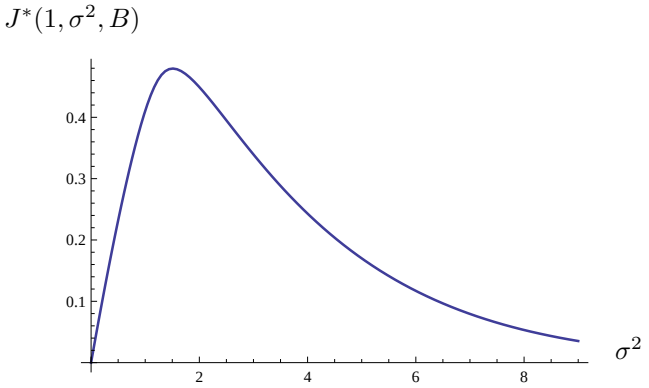

Fig. 3. $J^*(1, \sigma^2, B)$ against σ^2 where B is the Rademacher distribution.

Nonetheless, monotonicity in signal power holds for *Gaussian* input, an immediate consequence of Theorem 3:

Corollary 1.

(a) Noisy input costs more: For any distribution Q,

$$J^*(k^2, \sigma^2, P * Q) \ge J^*(k^2, \sigma^2, P) \tag{61}$$

(b) For Gaussian input, $\sigma^2 \mapsto J^*(k^2, \sigma^2)$ is increasing.

Proof. Observe that $P * Q$ is a location mixture of P. In view of the translation-invariance and concavity of J^* in P, (61) follows from applying Jensen's inequality. For (b), note that $J^*(k^2, \sigma^2) = J^*(k^2, 1, \mathcal{N}(0, \sigma^2))$. The desired monotonicity then follows from (61) and the infinite divisibility of Gaussian distribution. □

From the above proof we see that, monotonicity also holds for any stable input distribution [26] and any noise distribution (not necessarily Gaussian).

[4] As $\sigma^2 \to \infty$, $\sigma^2 \,\mathsf{mmse}(P, \sigma^2)$ converges to the *MMSE dimension* of P, which is zero for all discrete P [25, Theorem 4].

C. Optimal cost: Gaussian input

Theorem 5. *$\sigma^2 \mapsto J^*(k^2, \sigma^2)$ is increasing, subadditive and Lipschitz continuous, with*

$$0 \le \frac{\partial J^*}{\partial \sigma^2} \le \frac{k^2}{k^2 + 1}. \tag{62}$$

Proof. Since $\mathsf{mmse}(Q, \cdot)$ is decreasing,

$$\frac{J^*(k^2, \sigma^2)}{\sigma^2} = \min_Q \left\{ k^2 W_2^2(\mathcal{N}(0,1), Q) + \mathsf{mmse}(Q, \sigma^2) \right\} \tag{63}$$

is also decreasing in σ^2. This implies the desired subadditivity. Another consequence is

$$\frac{\partial J^*}{\partial \sigma^2} \le \frac{J^*}{\sigma^2} \le \left.\frac{\partial J^*}{\partial \sigma^2}\right|_{\sigma^2 = 0} = \frac{k^2}{k^2 + 1}, \tag{64}$$

where the last equality follows from (67) proved next. □

D. Weak-signal regime

By the continuity of MMSE [3, Proposition 7], for all Q with finite variance, $\mathsf{mmse}(Q, \sigma^2) = \mathsf{var} Q + o(1)$ as $\sigma^2 \to 0$. By Lemma 1(g) and (27), for any P,

$$\lim_{\sigma^2 \to 0} \frac{J^*(k^2, \sigma^2, P)}{\sigma^2} = \min_{P_Y} \left\{ k^2 W_2^2(P, Q) + \mathsf{var} Q \right\} \tag{65}$$

$$= \min_{\lambda \ge 0} k^2 (\sqrt{\mathsf{var} P} - \lambda)^2 + \lambda^2 \tag{66}$$

$$= \frac{k^2}{k^2 + 1} \mathsf{var} P, \tag{67}$$

attained by the *affine* controller

$$f(x) = \frac{k^2 \sqrt{\mathsf{var} P}}{k^2 + 1} (x - m_1(P)) + m_1(P). \tag{68}$$

E. Strong-signal regime

Fix k and let Q_σ^* be an optimizer of (27). Since $J^* \le 1$, we have

$$W_2^2(Q_\sigma^*, P) \le \frac{1}{k^2 \sigma^2} \tag{69}$$

which implies that $Q_\sigma^* \xrightarrow{W_2} P$ as $\sigma^2 \to \infty$. Therefore, the corresponding optimal controller f_σ^* also converges to the identity in $L^2(\mathbb{R}, P)$. However this does not imply almost sure convergence.

Note that as $\sigma \to \infty$, the asymptotically optimal affine controller converges the identity. This is equivalent to setting $Q = P$. However choosing $Q = P$ is *not* necessarily asymptotically optimal, even though the optimal output distribution Q_σ^* does converge to the input distribution P. This is because $Q_\sigma^* \xrightarrow{W_2} P$ does *not* imply that $\sigma^2 \mathsf{mmse}(Q_\sigma^*, \sigma^2) - \sigma^2 \mathsf{mmse}(P, \sigma^2) \to 0$. Indeed, for $P = \mathcal{N}(0,1)$ and all $k < 0.564$,

$$\lim_{\sigma \to \infty} J^*(k^2, \sigma^2) < 1 = \lim_{\sigma \to \infty} J_{\mathrm{a}}^*(k^2, \sigma^2). \tag{70}$$

To see this, let Q_σ to be the optimal m-point *uniform* quantized version of $\mathcal{N}(0,1)$ with $\sigma = \frac{2a}{\Delta_m}$, where Δ_m is the optimal step size and $a > 0$ is to be optimized later. By [27, Theorem 13], $\Delta_m = \frac{4\sqrt{\log m}}{m}(1 + o(1))$ and the optimal

mean-square quantization error is $D_m = \frac{1}{12}\Delta_m^2(1+o(1))$.[5] Therefore $W_2^2(Q_\sigma, \mathcal{N}(0,1)) \le D_m$. Let $Y_\sigma \sim Q_\sigma$ and $Z_\delta = \sigma Y_\sigma + N$. Define the following suboptimal estimator of N based on Z_σ: $f(z) = z - \sigma y(z/\sigma)$, where $y(x)$ is the closest atom of Y_σ to x. Such an estimator is exact whenever $|N| \le a$. Moreover, $N > a$ (resp. $N < -a$) implies that $-a < f(Z_\sigma) < 0$ (resp. $0 < f(Z_\sigma) < a$). Therefore

$$\begin{aligned}
\sigma^2\,\mathsf{mmse}(Q_\sigma, \sigma^2) &= \mathsf{mmse}(N|Z_\sigma) && (71)\\
&\le \mathbb{E}\left[(N - f(Z_\sigma))^2\right] && (72)\\
&\le 8\mathbb{E}\left[N^2 \mathbf{1}_{\{N>a\}}\right] && (73)\\
&= 8Q(a) + 8a\varphi(a) && (74)
\end{aligned}$$

Hence

$$\begin{aligned}
\lim_{\sigma\to\infty} J^*(k^2,\sigma^2) &\le \min_{a>0}\left\{\frac{4}{3}k^2a^2 + 8Q(a) + 8a\varphi(a)\right\} && (75)\\
&< 1 && (76)
\end{aligned}$$

for all $k < 0.564$.

VI. Concluding remarks

We gave a transport-theoretic formulation of Witsenhausen's counterexample. The Wasserstein metric (17) as well as the more general Monge-Kantorovich cost functional (14) are particularly relevant to decentralized stochastic decision problems with non-classical information structure where the the decision of the later-stage controller only depends on the *output distribution* of the controller in the earlier stage.

In addition to solving for the minimizer of (27) for a given P, there are several interesting open problems. Theorem 3 shows that $P \mapsto J^*(k^2,\sigma^2,P)$ is concave, upper semicontinuous and bounded. Therefore it makes sense to investigate the *worst-case* input distribution, for instance, in the following senses,

$$\max_{P:\,\mathsf{var}P\le 1} J^*(k^2,\sigma^2,P) \tag{77}$$

and

$$\max_{P:\,\mathrm{supp}(P)\subset[-A,A]} J^*(k^2,\sigma^2,P) \tag{78}$$

It is not clear whether the least favorable prior in (77) is Gaussian. As for (78), recall that under bounded support constraint, the least favorable prior for the mean-square error with Gaussian noise [28, p. 79] and the capacity-achieving distribution of Gaussian channel [29] are both finitely-supported. Using similar analyticity arguments, it might be possible to show that the maximizer of (78) is also finitely-supported.

We have shown that affine controllers are asymptotically optimal in the weak-signal regime ($\sigma \to 0$), but strictly suboptimal in the strong-signal regime ($\sigma \to \infty$) for all fixed $k < 0.564$. An open question is whether affine controllers are strictly suboptimal for *all* $\sigma > 0$ and $k > 0$. Since optimal affine controllers satisfy the variational equation (38), they are stationary points. Hence any proof of suboptimality based on local perturbation will fail.

Other open problems includes whether the minimizer of (27) is unique and symmetric. For symmetric P, choosing a symmetric Q decreases the W_2 distance but increases the MMSE.

[5] In fact the same conclusion holds for any generalized Gamma distribution [27, Section III.A].

References

[1] H. S. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM Journal on Control*, vol. 6, pp. 131 – 147, 1968.

[2] Y. Wu and S. Verdú, "Functional properties of MMSE," in *Proceedings of 2010 IEEE International Symposium on Information Theory*, Austin, TX, June 2010.

[3] D. Guo, Y. Wu, S. Shamai(Shitz), and S. Verdú, "Estimation in Gaussian Noise: Properties of the Minimum Mean-square Error," to appear in *IEEE Transactions on Information Theory*, 2011.

[4] R. Bansal and T. Başar, "Stochastic teams with nonclassical information revisited: When is an affine law optimal?" *IEEE Transactions on Automatic Control*, vol. 32, no. 6, pp. 554–559, Jun. 2002.

[5] S. Mitter and A. Sahai, "Information and control: Witsenhausen revisited," *Learning, control and hybrid systems*, pp. 281–293, 1999.

[6] M. Baglietto, T. Parisini, and R. Zoppoli, "Numerical solutions to the Witsenhausen counterexample by approximating networks," *IEEE Transactions on Automatic Control*, vol. 46, no. 9, pp. 1471–1477, Sep. 2002.

[7] J. Lee, E. Lau, and Y. Ho, "The Witsenhausen counterexample: A hierarchical search approach for nonconvex optimization problems," *IEEE Transactions on Automatic Control*, vol. 46, no. 3, pp. 382–397, 2002.

[8] N. Li, J. R. Marden, and J. S. Shamma, "Learning approaches to the Witsenhausen counterexample from a view of potential games," in *Joint 48th IEEE Conference on Decision and Control and 28th Chinese Control Conference*, Dec. 2009, pp. 157–162.

[9] P. Grover and A. Sahai, "Witsenhausen's counterexample as Assisted Interference Suppression," *International Journal of Systems, Control and Communications*, vol. 2, no. 1, pp. 197–237, 2010.

[10] P. Grover, S. Park, and A. Sahai, "The finite-dimensional Witsenhausen counterexample," *Submitted to IEEE Transactions on Automatic Control*, 2010.

[11] Y. C. Ho, "Review of the Witsenhausen problem," in *Proceedings of the 47th IEEE Conference on Decision and Control*, Dec. 2008, pp. 1614–1619.

[12] T. Başar, "Variations on the theme of the Witsenhausen counterexample," in *Proceedings of the 47th IEEE Conference on Decision and Control*, Dec. 2008, pp. 1614–1619.

[13] C. Villani, *Optimal Transport: Old and New*. Berlin: Springer Verlag, 2008.

[14] E. Çinlar, *Probability and Stochastics*. New York: Springer, 2011.

[15] S. T. Rachev and L. Rüschendorf, *Mass Transportation Problems: Vol. I: Theory*. Springer-Verlag, 1998.

[16] G. Dall'Aglio, "Sugli estremi dei momenti delle funzioni di ripartizione doppia," *Ann. Scuola Norm. Sup. Pisa*, vol. 10, pp. 35–74, 1956.

[17] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*, 2nd ed. Basel, Switzerland: Birkhäuser, 2008.

[18] M. H. DeGroot, *Optimal Statistical Decisions*. New York, NY: McGraw-Hill, 1970.

[19] S. Yüksel and T. Linder, "Optimization and Convergence of Observation Channels in Stochastic Control," 2010, submitted to *SIAM Journal on Control and Optimization*.

[20] R. Jordan, D. Kinderlehrer, and F. Otto, "The variational formulation of the Fokker-Planck equation," *SIAM journal on mathematical analysis*, vol. 29, no. 1, pp. 1–17, 1998.

[21] E. A. Carlen and W. Gangbo, "Constrained steepest descent in the 2-Wasserstein metric," *Annals of mathematics*, pp. 807–846, 2003.

[22] A. Tudorascu, "On the Jordan–Kinderlehrer–Otto variational scheme and constrained optimization in the Wasserstein metric," *Calculus of Variations and Partial Differential Equations*, vol. 32, no. 2, pp. 155–173, 2008.

[23] J. Dieudonné, *Foundations of Modern Analysis*. New York, NY: Academic Press, 1969.

[24] G. Birkhoff and S. Mac Lane, *Algebra*. New York, NY: Chelsea, 1988.

[25] Y. Wu and S. Verdú, "MMSE Dimension," in *Proceedings of 2010 IEEE International Symposium on Information Theory*, Austin, TX, June 2010.

[26] V. M. Zolotarev, *One-dimensional Stable Distributions*. Providence, RI: American Mathematical Society, 1986.

[27] D. Hui and D. Neuhoff, "Asymptotic analysis of optimal fixed-rate uniform scalar quantization," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 957–977, Mar. 2002.

[28] I. Johnstone, "Function Estimation and Gaussian Sequence Models," unpublished lecture notes. [Online]. Available: www-stat.stanford.edu/ imj/baseb.pdf

[29] J. G. Smith, "The information capacity of amplitude and variance-constrained scalar gaussian channels," *Information and Control*, vol. 18, pp. 203 – 219, 1971.