Minimum-Information LQG Control
Supplementary Material

Roy Fox[†] and Naftali Tishby[†]

Appendix I
Perfectly matched channel

In this appendix we construct a channel that is perfectly matched to the sequential source code derived in Theorem 1, in Part I of this paper [1, Section III-B]. Recall that in a perfectly matched source-channel pair the optimal source coding and the optimal channel coding can be implemented jointly for single letters, without requiring longer blocks. This allows us to use them in a perception-action cycle, where we cannot accumulate a block of inputs before emitting an output.

The main results of [2], applied to our setting, can be summarized as follows. We wish to find a memoryless channel into which we can input an encoding $w_t = g(\hat{x}_{y_t})$, such that $\hat{x}_{u_t} = h(\hat{w}_t)$ can be decoded from the channel output $\hat{w}_t$. Suppose that we are concerned with the power needed to transmit w_t and thus the input cost is $w_t^\intercal w_t$. Then the source $\hat{x}_{y_t}$ and the channel $w_t \to \hat{w}_t$ are perfectly matched if there exist an encoder and a decoder such that

1) The Kullback-Leibler divergence $\mathbb{D}[f(\hat{w}_t|w_t)\|f(\hat{w}_t)]$ between the conditional and marginal densities of $\hat{w}_t$, as a function of w_t, equals $c_1 w_t^\intercal w_t + c_2$, for some constants $c_1 \geq 0$ and c_2; and
2) $f(\hat{x}_{u_t}|\hat{x}_{y_t})$ satisfies the conditions in Theorem 1.

To meet these conditions, we can choose the channel, the encoder and the decoder to have

$$\begin{aligned} w_t &= D^{1/2} V^\intercal \Sigma_{\hat{x}_y}^{\dagger/2} \hat{x}_{y_t} \\ \hat{w}_t &= w_t + \upsilon_t; \qquad\qquad \upsilon_t \sim \mathcal{N}(0, I - D) \\ \hat{x}_{u_t} &= \Sigma_{\hat{x}_y}^{1/2} V D^{1/2} \hat{w}_t, \end{aligned}$$

with D and V as in Theorem 1. Then

$$\begin{aligned} \Sigma_w &= D \\ \Sigma_{\hat{w}} &= I \\ \Sigma_{\hat{x}_u} &= \Sigma_{\hat{x}_y}^{1/2} V D V^\intercal \Sigma_{\hat{x}_y}^{1/2} = \Sigma_{\hat{x}_u;\hat{x}_y}, \end{aligned}$$

and it can be verified that

$$\mathbb{D}[f(\hat{w}_t|w_t)\|f(\hat{w}_t)] = \tfrac{1}{2} w_t^\intercal \Sigma_{\hat{w}}^{-1} w_t + \text{const},$$

as required.

The capacity of the additive Gaussian noise channel with noise covariance $I - D$, under the appropriate expected power constraint, is indeed achieved by a Gaussian input with covariance D and is equal to the information rate in Theorem 1. As shown in [2], this means that constraining the expected power Σ_w is equivalent to constraining the information rate $\mathbb{I}[\hat{x}_{y_t}; \hat{x}_{u_t}]$.

Note, however, that the matched channel noise covariance depends on the constraint, through the solution in Theorem 1. Moreover, this result is not applicable when the best channel available to the designer of the controller is not the matched channel above, in which case both the channel and the sequential source coding generally need to be adapted.

Appendix II
Proof of Lemma 1 of Part I

In this appendix we restate and prove Lemma 1 of Part I [1, Section IV-A].

Lemma 1: Let x and $\hat{x}$ be 0-mean jointly Gaussian random variables. The following properties are equivalent:

1) There exists a random variable u, jointly Gaussian with x, such that $\hat{x}(u) = \arg\min_{\hat{x}} \mathbb{E}[\|\hat{x} - x\|^2 | u] = \mathbb{E}[x|u]$.
2) $\Sigma_{\hat{x};x} = \Sigma_{\hat{x}}$.
3) $\Sigma_{x|\hat{x}} = \Sigma_x - \Sigma_{\hat{x}}$, where $\Sigma_{x|\hat{x}}$ is the conditional covariance matrix of x given $\hat{x}$, implying $\Sigma_x \succeq \Sigma_{\hat{x}}$.
4) $\hat{x} = \mathbb{E}[x|\hat{x}]$.

Such $\hat{x}$ is called a minimum mean square error (MMSE) estimator (of u) for x.

Proof: (1 $\implies$ 2) Assume without loss of generality that u has mean 0. Then

$$\hat{x} = \Sigma_{x;u} \Sigma_u^\dagger u,$$

implying

$$\Sigma_{\hat{x};x} = \Sigma_{x;u} \Sigma_u^\dagger \Sigma_{u;x} = \Sigma_{\hat{x}} .$$

(2 $\implies$ 3)

$$\Sigma_{x|\hat{x}} = \Sigma_x - \Sigma_{x;\hat{x}} \Sigma_{\hat{x}}^\dagger \Sigma_{\hat{x};x} = \Sigma_x - \Sigma_{\hat{x}} .$$

(3 $\implies$ 4) Since x and $\hat{x}$ are 0-mean and jointly Gaussian, we can write for some T

$$x = T\hat{x} + \xi; \qquad\qquad \xi \sim \mathcal{N}(0, \Sigma_{x|\hat{x}}),$$

implying

$$\Sigma_x = T \Sigma_{\hat{x}} T^\intercal + \Sigma_x - \Sigma_{\hat{x}},$$

thus without loss of generality $T = I$.

[†]School of Computer Science and Engineering, The Hebrew University, {royf,tishby}@cs.huji.ac.il

*This work was supported by the DARPA MSEE Program, the Gatsby Charitable Foundation, the Israel Science Foundation and the Intel ICRI-CI Institute

(4 $\implies$ 1) Taking $u = \hat{x}$, we have

$$\begin{aligned}&\arg\min_{\hat{x}'} \mathbb{E}[\|\hat{x}' - x\|^2 | u] \\ &\quad = \arg\min_{\hat{x}'} (\hat{x}'^{\mathsf{T}} \hat{x}' - 2\hat{x}'^{\mathsf{T}} \mathbb{E}[x|u]) + \mathbb{E}[x^{\mathsf{T}} x | u],\end{aligned}$$

which is optimized by $\hat{x}' = \mathbb{E}[x|u]$. □

Appendix III
Proof of Lemma 2 of Part I

In this appendix we restate and prove Lemma 2 of Part I [1, Section IV-A].

Lemma 2: The bounded memoryless LTI controller optimization problem (Problem 1) is solved by a control law of the form

$$\hat{x}_{y_t} = K y_t \tag{5a}$$
$$\hat{x}_{u_t} = W \hat{x}_{y_t} + \omega_t; \qquad \omega_t \sim \mathcal{N}(0, \Sigma_\omega) \tag{5b}$$
$$u_t = L \hat{x}_{u_t}, \tag{5c}$$

where $W \in \mathbb{R}^{n \times n}$, $\Sigma_\omega \in \mathbb{R}^{n \times n}$, $L \in \mathbb{R}^{\ell \times n}$, ω_t is independent of y_t, $\hat{x}_{u_t}$ is a MMSE estimator for $\hat{x}_{y_t}$ and

$$\mathbb{I}[y_t; u_t] = \mathbb{I}[\hat{x}_{y_t}; \hat{x}_{u_t}]. \tag{6}$$

Proof: Consider a LTI controller π of the form

$$u_t = H y_t + \eta_t; \qquad \eta_t \sim \mathcal{N}(0, \Sigma_\eta), \tag{III.1}$$

satisfying the Markov network

$$\begin{array}{ccccc} x_t & — & y_t & — & u_t \\ & & | & & | \\ & & \hat{x}_{y_t} & & \hat{x}_{u_t}. \end{array}$$

We now construct a controller π' with control law u'_t based on the estimator $\hat{x}'_{u_t}$ by defining the Markov chain

$$x_t \ — \ y_t \ — \ \hat{x}_{y_t} \ — \ u''_t \ — \ \hat{x}'_{u_t} \ — \ u'_t$$

such that each consecutive pair of variables has the same joint distribution as their unprimed namesakes. Since $\hat{x}_{y_t}$ is a sufficient statistic of y_t for x_t, we have the Markov chain x_t — $\hat{x}_{y_t}$ — y_t — u_t, implying that u''_t has the same joint distribution with x_t as u_t does. Likewise, $\hat{x}'_{u_t}$ has the same joint distribution with x_t as $\hat{x}_{u_t}$ does. Since $\hat{x}_{u_t}$ is a sufficient statistic of u_t for x_t, we have that u'_t also has the same joint distribution with x_t as u_t does.

Thus the controller π' induces the same stochastic process $\{x_t, u'_t\}$ and the same external cost. Note that u'_t may not have the same joint distribution with y_t as u_t does and due to the data-processing inequality [3]

$$\begin{aligned}\mathbb{I}[y_t; u_t] &\geq \mathbb{I}[\hat{x}_{y_t}; u_t] = \mathbb{I}[\hat{x}_{y_t}; u''_t] \\ &\geq \mathbb{I}[\hat{x}_{y_t}; \hat{x}'_{u_t}] \geq \mathbb{I}[y_t; u'_t].\end{aligned}$$

Therefore π' performs at least as well as π and equally well when π is optimal, proving (6).

$\hat{x}'_{u_t}$ is a MMSE estimator for $\hat{x}_{y_t}$ since

$$\begin{aligned}\mathbb{E}[\hat{x}_{y_t} | \hat{x}'_{u_t}] &= \mathbb{E}[\mathbb{E}[x_t | y_t] | \hat{x}_{u'_t}] \\ &= \mathbb{E}[x_t | \hat{x}'_{u_t}] = \hat{x}'_{u_t},\end{aligned}$$

where the second equality follows from x_t — y_t — $\hat{x}'_{u_t}$.

Finally, it may not be clear from the above analysis that u'_t is optimally deterministic in $\hat{x}'_{u_t}$. If u_t has covariance Σ_ν given $\hat{x}'_{u_t}$, the Lagrangian of the optimization problem ((9) in Part I) depends on Σ_ν only through the terms

$$\tfrac{1}{2}(\operatorname{tr}(R \Sigma_\nu) + \operatorname{tr}(S B \Sigma_\nu B^{\mathsf{T}})).$$

Since $R + B^{\mathsf{T}} S B \succeq 0$ is positive semidefinite, we can take $\Sigma_\nu = 0$ without loss of performance, recovering the structure (5). Intuitively, the argument is that any noise added to u'_t, beyond $\hat{x}'_{u_t}$, is not helpful in compressing x_t and can only increase the external cost without saving any communication cost.

In the other direction, let u_t satisfy the form of Lemma 2. We can rewrite u_t in the form (III.1), with

$$\begin{aligned}H &= LWK \\ \Sigma_\eta &= L \Sigma_\omega L^{\mathsf{T}}.\end{aligned}$$ □

Appendix IV
Proof of Theorem 1 of Part I

In this appendix we restate and prove Theorem 1 of Part I [1, Section IV-A], which relies on the following Lagrangian developed there.

$$\begin{aligned}\mathcal{F}_{\Sigma_x, \Sigma_{\hat{x}_u}, L, S; \beta} = \tfrac{1}{2}(&\beta^{-1}(\log|\Sigma_{\hat{x}_y}|_\dagger - \log|\Sigma_{\hat{x}_y|\hat{x}_u}|_\dagger) \\ &+ \operatorname{tr}(Q \Sigma_x) + \operatorname{tr}(RL \Sigma_{\hat{x}_u} L^{\mathsf{T}}) \\ &+ \operatorname{tr}(S((A+BL)\Sigma_{\hat{x}_u}(A+BL)^{\mathsf{T}} \\ &\quad + A \Sigma_{x|\hat{x}_u} A^{\mathsf{T}} + \Sigma_\xi - \Sigma_x))).\end{aligned} \tag{9}$$

Theorem 1: Given β, the Lagrangian (9) is minimized by a controller satisfying the forward equations

$$\begin{aligned}\Sigma_x &= (A+BL)\Sigma_{\hat{x}_u}(A+BL)^{\mathsf{T}} \\ &\quad + A \Sigma_{x|\hat{x}_u} A^{\mathsf{T}} + \Sigma_\xi \qquad (10a)\\ \Sigma_y &= C \Sigma_x C^{\mathsf{T}} + \Sigma_\epsilon \qquad (10b)\\ K &= \Sigma_x C^{\mathsf{T}} \Sigma_y^\dagger \qquad (10c)\\ \Sigma_{\hat{x}_y} &= K \Sigma_y K^{\mathsf{T}}, \qquad (10d)\end{aligned}$$

the backward equations

$$\begin{aligned}M &= \beta^{-1} C^{\mathsf{T}} K^{\mathsf{T}} (\Sigma^\dagger_{\hat{x}_y|\hat{x}_u} - \Sigma^\dagger_{\hat{x}_y}) K C \qquad (10e)\\ S &= Q + A^{\mathsf{T}} S A - M, \qquad (10f)\\ L &= -(R + B^{\mathsf{T}} S B)^\dagger B^{\mathsf{T}} S A \qquad (10g)\\ N &= L^{\mathsf{T}}(R + B^{\mathsf{T}} S B) L \qquad (10h)\end{aligned}$$

and the control-based estimator covariance

$$\Sigma_{\hat{x}_u} = \Sigma^{1/2}_{\hat{x}_y} V D V^{\mathsf{T}} \Sigma^{1/2}_{\hat{x}_y}, \tag{10i}$$

the latter determined by the eigenvalue decomposition (EVD)

$$V \Lambda V^{\mathsf{T}} = \Sigma^{1/2}_{\hat{x}_y} N \Sigma^{1/2}_{\hat{x}_y} \tag{10j}$$

having V orthogonal with $n - \operatorname{rank}(\Sigma_{\hat{x}_y})$ columns spanning the kernel of $\Sigma_{\hat{x}_y}$ and $\Lambda = \operatorname{diag}\{\lambda_i\}$ and by the active mode coefficient matrix

$$D = \operatorname{diag}\left\{ \begin{array}{ll} 1 - \beta^{-1}\lambda_i^{-1} & \lambda_i > \beta^{-1} \\ 0 & \lambda_i \leq \beta^{-1} \end{array} \right\}. \tag{10k}$$

Proof: The minimum of the Lagrangian (9) must satisfy the first-order optimality conditions, i.e. that the gradient with respect to each parameter is 0 at the optimum. We start by differentiating $\mathcal{F}$ by the feedback gain L

$$\partial_L \mathcal{F}_{\Sigma_x,\Sigma_{\hat{x}_u},L,S;\beta} = RL\,\Sigma_{\hat{x}_u} + B^\intercal S(A+BL)\,\Sigma_{\hat{x}_u} = 0,$$

which we rewrite as

$$(R+B^\intercal SB)L\,\Sigma_{\hat{x}_u} = -B^\intercal SA\,\Sigma_{\hat{x}_u}\,.$$

As this equation shows, L is underdetermined in the kernel of $\Sigma_{\hat{x}_u}$, since these modes are always 0 in $\hat{x}_{u_t}$ and have no effect on u_t. L is also underdetermined in the kernel of $R+B^\intercal SB$, since these modes have no cost (immediate or future) and can be controlled in any way without affecting the solution's performance. Thus without loss of performance we can take

$$L = -(R+B^\intercal SB)^\dagger B^\intercal SA.$$

We substitute this solution back into the Lagrangian, to get

$$\begin{aligned}\mathcal{F}_{\Sigma_x,\Sigma_{\hat{x}_u},S;\beta} = \tfrac{1}{2}(\beta^{-1}(\log|\,\Sigma_{\hat{x}_y}\,|_\dagger - \log|\,\Sigma_{\hat{x}_y|\hat{x}_u}\,|_\dagger) \\ + \operatorname{tr}(M\,\Sigma_x) - \operatorname{tr}(N\,\Sigma_{\hat{x}_u}) + \operatorname{tr}(S\,\Sigma_\xi)),\end{aligned} \tag{IV.1}$$

with

$$\begin{aligned} M &= Q + A^\intercal SA - S \\ N &= L^\intercal(R+B^\intercal SB)L \\ &= A^\intercal SB(R+B^\intercal SB)^\dagger B^\intercal SA. \end{aligned}$$

The problem of optimizing over $\Sigma_{\hat{x}_u}$ given the other parameters can now be written, up to constants, as the semidefinite program (SDP)

$$\begin{aligned} \max_{\Sigma_{\hat{x}_u}} \quad & \log|\,\Sigma_{\hat{x}_y} - \Sigma_{\hat{x}_u}\,|_\dagger + \beta\operatorname{tr}(N\,\Sigma_{\hat{x}_u}) \\ \text{s.t.} \quad & 0 \preceq \Sigma_{\hat{x}_u} \preceq \Sigma_{\hat{x}_y}\,. \end{aligned}$$

By Lemma V.1 in Appendix V, the optimum is achieved when $\Sigma_{\hat{x}_u}$ satisfies (10i)–(10k).

Finally, with $P = \Sigma_{\hat{x}_y}\,\Sigma^\dagger_{\hat{x}_y}$ the projection onto the support of $\hat{x}_{y_t}$ and since the range of $\Sigma_{\hat{x}_u}$ is contained in that subspace, we have

$$\begin{aligned}&\partial_{(\Sigma_x)_{i,j}}(\log|\,\Sigma_{\hat{x}_y}\,|_\dagger - \log|\,\Sigma_{\hat{x}_y|\hat{x}_u}\,|_\dagger) \\ &\quad= -\,\partial_{(\Sigma_x)_{i,j}}\log|P - \Sigma_{\hat{x}_u}\,\Sigma^\dagger_{\hat{x}_y}\,|_\dagger \\ &\quad= -\,\partial_{(\Sigma_x)_{i,j}}\log|I - \Sigma_{\hat{x}_u}(P\,\Sigma_{\hat{x}_y}\,P)^\dagger| \\ &\quad= \operatorname{tr}((I-\Sigma_{\hat{x}_u}\,\Sigma^\dagger_{\hat{x}_y})^{-1}\,\Sigma_{\hat{x}_u}\,\partial_{(\Sigma_x)_{i,j}}(P\,\Sigma_{\hat{x}_y}\,P)^\dagger).\end{aligned}$$

The purpose of introducing P is to notice that even if the range of $\Sigma_{\hat{x}_y}$ is increased, this has no effect on the Lagrangian, because these modes are orthogonal to the range of $\Sigma_{\hat{x}_u}$. This allows us to treat P as constant, so that the range of $P\,\Sigma_{\hat{x}_y}\,P$ is constant in a neighborhood of the solution, and the derivative of the pseudoinverse is simplified in this case to

$$\begin{aligned}\partial_{(\Sigma_x)_{i,j}}(P\,\Sigma_{\hat{x}_y}\,P)^\dagger &= -\,\Sigma^\dagger_{\hat{x}_y}(\partial_{(\Sigma_x)_{i,j}}\,\Sigma_{\hat{x}_y})\,\Sigma^\dagger_{\hat{x}_y} \\ &= -\,\Sigma^\dagger_{\hat{x}_y}\,KCJ_{i,j}C^\intercal K^\intercal\,\Sigma^\dagger_{\hat{x}_y},\end{aligned}$$

with $J_{i,j}$ the matrix with 1 in position (i,j) and 0 elsewhere. This yields

$$\begin{aligned}&\partial_{\Sigma_x}\mathcal{F}_{\Sigma_x,\Sigma_{\hat{x}_u},S;\beta} \\ &\quad= \tfrac{1}{2}(M - \beta^{-1}C^\intercal K^\intercal\,\Sigma^\dagger_{\hat{x}_y}(I-\Sigma_{\hat{x}_u}\,\Sigma^\dagger_{\hat{x}_y})^{-1}\,\Sigma_{\hat{x}_u}\,\Sigma^\dagger_{\hat{x}_y}\,KC) \\ &\quad= \tfrac{1}{2}(M - \beta^{-1}C^\intercal K^\intercal\,\Sigma^\dagger_{\hat{x}_y}((I-\Sigma_{\hat{x}_u}\,\Sigma^\dagger_{\hat{x}_y})^{-1} - I)KC) \\ &\quad= \tfrac{1}{2}(M - \beta^{-1}C^\intercal K^\intercal(\Sigma^\dagger_{\hat{x}_y|\hat{x}_u} - \Sigma^\dagger_{\hat{x}_y})KC) = 0,\end{aligned}$$

implying (10e). □

Appendix V
Semidefinite program solution

In this appendix we state and prove the following solution to our SDP problem.

Lemma V.1: The semidefinite program

$$\begin{aligned} \max_{X\in\mathbb{S}^n_+} \quad & \log|M_1 - X|_\dagger + \operatorname{tr}(M_2X) \\ \text{s.t.} \quad & X \preceq M_1, \end{aligned}$$

with $M_1, M_2 \succeq 0$, is optimized by

$$X = M_1^{1/2}VDV^\intercal M_1^{1/2},$$

with the eigenvalue decomposition (EVD)

$$V\Lambda V^\intercal = M_1^{1/2}M_2M_1^{1/2},$$

such that V is orthogonal with $n - \operatorname{rank}(M_1)$ columns spanning the kernel of M_1 and $\Lambda = \operatorname{diag}\{\lambda_i\}$ and with

$$D = \operatorname{diag}\left\{\begin{array}{ll} 1-\lambda_i^{-1} & \lambda_i > 1 \\ 0 & \lambda_i \le 1 \end{array}\right\}.$$

Proof: Let the EVD of M_1 be

$$U\Psi U^\intercal = M_1,$$

with U orthogonal and Ψ diagonal, having

$$\Psi = \begin{bmatrix}\Psi_+ & 0 \\ 0 & 0_{(n-m)\times(n-m)}\end{bmatrix},$$

with $m = \operatorname{rank}(M_1)$. Let

$$\Psi^\ddagger = \Psi^\dagger + I - \Psi^\dagger\Psi = \begin{bmatrix}\Psi_+^{-1} & 0 \\ 0 & I\end{bmatrix}.$$

By changing the variable to

$$Y = \Psi^{\ddagger/2}U^\intercal XU\Psi^{\ddagger/2},$$

the constraint of the SDP becomes

$$Y \preceq I_{m,n} = \begin{bmatrix}I_{m\times m} & 0 \\ 0 & 0_{(n-m)\times(n-m)}\end{bmatrix}.$$

Y must therefore be 0 outside the upper-left $m\times m$ block, and the SDP is equivalent, up to constants, to

$$\begin{aligned} \max_{Y\in\mathbb{S}^n_+} \quad & \log|I_{m,n} - Y|_\dagger + \operatorname{tr}(\Psi^{1/2}U^\intercal M_2U\Psi^{1/2}Y) \\ \text{s.t.} \quad & Y \preceq I_{m,n}. \end{aligned}$$

Let the EVD of the linear coefficient be

$$\bar{V}\bar{\Lambda}\bar{V}^\intercal = \Psi^{1/2}U^\intercal M_2U\Psi^{1/2},$$

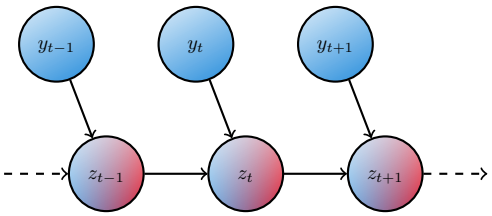

Fig. VI.1. Bayesian network of online inference from a sequence of independent observations

with

$$\bar{V} = \begin{bmatrix} \bar{V}_+ & 0 \\ 0 & I_{(n-m)\times(n-m)} \end{bmatrix}$$

orthogonal and preserving the kernel of Ψ and $\Lambda = \operatorname{diag}\{\lambda_i\}$. We can again change the variable to

$$D = \bar{V}^\intercal Y \bar{V},$$

to get

$$\begin{aligned} \max_{D \in \mathbb{S}^n_+} \quad & \log|I_{m,n} - D|_\dagger + \operatorname{tr}(\Lambda D) \\ \text{s.t.} \quad & D \preceq I_{m,n}, \end{aligned}$$

which can easily be solved using Hadamard's inequality [3], to find

$$D = \operatorname{diag}\left\{ \begin{array}{ll} 1 - \lambda_i^{-1} & \lambda_i > 1 \\ 0 & \lambda_i \leq 1 \end{array} \right\}.$$

Finally, the lemma follows by unmaking the variable changes and taking

$$V = U\bar{V}. \qquad \square$$

Appendix VI
Properties of the retentive directed information

In this appendix we show how the retentive directed information (Definition 6 of Part II [4, Section III-A]) relates to the multi-information of Bayesian networks [5].

Consider the Bayesian network in Figure VI.1, which describes the process of online inference from a sequence of independent observations. The multi-information of this network, for horizon T, is equal to the retentive directed information

$$\begin{aligned} \mathbb{I}[y^T, z^T] &= \mathbb{E}\left[\log \frac{f(y^T, z^T)}{\prod_{t=1}^T f(y_t) f(z_t)}\right] \\ &= \sum_{t=1}^T \mathbb{E}\left[\log \frac{f(z_t|z^{t-1}, y^t)}{f(z_t)}\right] = \mathbb{I}[y^T \twoheadrightarrow z^T]. \end{aligned}$$

An important property of the directed information is that the mutual information between two sequences can be decomposed into the sum of directed information in both directions [6]

$$\mathbb{I}[x^T; z^T] = \mathbb{I}[x^T \to z^T] + \mathbb{I}[z^T \to x^T].$$

Interestingly, retentive directed information extends this property to the retentive control process (Figure 1 in Part II). This process can be thought of as consisting of four phases: observation, inference, control and state transition. Its multi-information can accordingly be decomposed [7] into the sum

$$\begin{aligned} \mathbb{I}[x^T, y^T, z^T, u^T] = \; & \mathbb{I}[x^T \to y^T] + \mathbb{I}[y^T \twoheadrightarrow z^T] \\ & + \mathbb{I}[z^T \to u^T] + \mathbb{I}[u^T \twoheadrightarrow x^T]. \end{aligned}$$

Appendix VII
Structure of the optimal retentive controller

In this appendix we derive the structure of the optimal retentive controller summarized in Part II [4, Section III-C].

For the structured feedback gain L we find using the Schur complement that

$$\begin{aligned} (R + B^\intercal S B)^\dagger &= \begin{bmatrix} R_u + B_{x;u}^\intercal S_x B_{x;u} & B_{x;u}^\intercal S_{x;m} \\ S_{m;x} B_{x;u} & S_m \end{bmatrix}^\dagger \\ &= \begin{bmatrix} S_{u|m}^\dagger & -S_{u|m}^\dagger B_{x;u}^\intercal S_{x;m} S_m^\dagger \\ -S_m^\dagger S_{m;x} B_{x;u} S_{u|m}^\dagger & S_{m|u}^\dagger \end{bmatrix}, \end{aligned}$$

with

$$S_{m|u}^\dagger = S_m^\dagger + S_m^\dagger S_{m;x} B_{x;u} S_{u|m}^\dagger B_{x;u}^\intercal S_{x;m} S_m^\dagger,$$

and so

$$\begin{aligned} L &= -(R + B^\intercal S B)^\dagger B^\intercal S A \\ &= -(R + B^\intercal S B)^\dagger \begin{bmatrix} B_{x;u}^\intercal S_x A_x & 0 \\ S_{m;x} A_x & 0 \end{bmatrix} \\ &= -\begin{bmatrix} S_{u|m}^\dagger B_{x;u}^\intercal S_{x|m} A_x & 0 \\ S_m^\dagger S_{m;x} (I - B_{x;u} S_{u|m}^\dagger B_{x;u}^\intercal S_{x|m}) A_x & 0 \end{bmatrix} \\ &= \begin{bmatrix} L_{u;x|m} & 0 \\ -S_m^\dagger S_{m;x} (A_x + B_{x;u} L_{u;x|m}) & 0 \end{bmatrix}, \end{aligned}$$

with

$$L_{u;x|m} = -S_{u|m}^\dagger B_{x;u}^\intercal S_{x|m} A_x.$$

We also have

$$\begin{aligned} N &= L^\intercal (R + B^\intercal S B) L \\ &= A^\intercal S B (R + B^\intercal S B)^\dagger B^\intercal S A = \begin{bmatrix} N_{x|m} & 0 \\ 0 & 0 \end{bmatrix} \\ N_{x|m} &= \begin{bmatrix} B_{x;u}^\intercal S_x A_x \\ S_{m;x} A_x \end{bmatrix}^\intercal L \begin{bmatrix} I \\ 0 \end{bmatrix} \\ &= A_x^\intercal (S_x - S_{x|m} + S_{x|m} B_{x;u} S_{u|m}^\dagger B_{x;u}^\intercal S_{x|m}) A_x. \end{aligned}$$

Dually, for the structured Kalman gain K we find that

$$\begin{aligned} \Sigma_{\tilde{y}}^\dagger &= \begin{bmatrix} \Sigma_y & \Sigma_{y;m} \\ \Sigma_{m;y} & \Sigma_m \end{bmatrix}^\dagger \\ &= \begin{bmatrix} \Sigma_{y|m}^\dagger & -\Sigma_{y|m}^\dagger \Sigma_{y;m} \Sigma_m^\dagger \\ -\Sigma_m^\dagger \Sigma_{m;y} \Sigma_{y|m}^\dagger & \Sigma_m^\dagger + \Sigma_m^\dagger \Sigma_{m;y} \Sigma_{y|m}^\dagger \Sigma_{y;m} \Sigma_m^\dagger \end{bmatrix}, \end{aligned}$$

and so

$$\begin{aligned} K &= \Sigma_x C^\intercal \Sigma_{\tilde{y}}^\dagger \\ &= \begin{bmatrix} \Sigma_x C_{y;x}^\intercal & \Sigma_{x;m} \end{bmatrix} \begin{bmatrix} \Sigma_y & \Sigma_{y;m} \\ \Sigma_{m;y} & \Sigma_m \end{bmatrix}^\dagger \\ &= \begin{bmatrix} K_{x;y|m} & (I - K_{x;y|m} C_{y;x}) \Sigma_{x;m} \Sigma_m^\dagger \end{bmatrix}, \end{aligned}$$

with

$$K_{x;y|m} = \Sigma_{x|m}\, C_{y;x}^{\intercal}\, \Sigma_{y|m}^{\dagger}\,.$$

Now constraining the controller to be MMSE, we have the structure

$$\Sigma_{\tilde{x}} = \begin{bmatrix} \Sigma_{x|m} + \Sigma_m & \Sigma_m \\ \Sigma_m & \Sigma_m \end{bmatrix}$$
$$K = \begin{bmatrix} K_{x;y|m} & I - K_{x;y|m} C_{y;x} \end{bmatrix},$$

which we employ in differentiating $\mathcal{F}$ (IV.1), to get

$$\partial_{\Sigma_{x|m}} \mathcal{F}_{\Sigma_{x|m},\Sigma_m,\Sigma_{\hat{x}_{\tilde{u}}},S;\beta} = \begin{bmatrix} I \\ 0 \end{bmatrix}^{\intercal} \partial_{\Sigma_{\tilde{x}}} \mathcal{F}_{\Sigma_{\tilde{x}},\Sigma_{\hat{x}_{\tilde{u}}},S;\beta} \begin{bmatrix} I \\ 0 \end{bmatrix}$$
$$= \tfrac{1}{2} \begin{bmatrix} I \\ 0 \end{bmatrix}^{\intercal} (M - \beta^{-1} C^{\intercal} K^{\intercal} Z K C) \begin{bmatrix} I \\ 0 \end{bmatrix}$$
$$= \tfrac{1}{2} \left(\begin{bmatrix} I \\ 0 \end{bmatrix}^{\intercal} M \begin{bmatrix} I \\ 0 \end{bmatrix} - \beta^{-1} C_{y;x}^{\intercal} K_{x;y|m}^{\intercal} Z K_{x;y|m} C_{y;x} \right) = 0$$
$$\partial_{\Sigma_m} \mathcal{F}_{\Sigma_{x|m},\Sigma_m,\Sigma_{\hat{x}_{\tilde{u}}},S;\beta} = \begin{bmatrix} I \\ I \end{bmatrix}^{\intercal} \partial_{\Sigma_{\tilde{x}}} \mathcal{F}_{\Sigma_{\tilde{x}},\Sigma_{\hat{x}_{\tilde{u}}},S;\beta} \begin{bmatrix} I \\ I \end{bmatrix}$$
$$= \tfrac{1}{2} \begin{bmatrix} I \\ I \end{bmatrix}^{\intercal} (M - \beta^{-1} C^{\intercal} K^{\intercal} Z K C) \begin{bmatrix} I \\ I \end{bmatrix}$$
$$= \tfrac{1}{2} \left(\begin{bmatrix} I \\ I \end{bmatrix}^{\intercal} M \begin{bmatrix} I \\ I \end{bmatrix} - \beta^{-1} Z \right) = 0,$$

with

$$Z = \Sigma_{\hat{x}_{\tilde{y}}|\hat{x}_{\tilde{u}}}^{\dagger} - \Sigma_{\hat{x}_{\tilde{y}}}^{\dagger}\,.$$

This leaves M overparameterized and we can choose to give it the structure

$$M = \begin{bmatrix} M_{x|m} + M_m & -M_m \\ -M_m & M_m \end{bmatrix}$$

with

$$M_{x|m} = \beta^{-1} Z$$
$$M_m = \beta^{-1} (C_{y;x}^{\intercal} K_{x;y|m}^{\intercal} Z K_{x;y|m} C_{y;x} - Z).$$

References

[1] R. Fox and N. Tishby, "Minimum-information LQG control — Part I: Memoryless controllers," in *Proceedings of the 55th IEEE Conference on Decision and Control (CDC)*, 2016. [Online]. Available: https://arxiv.org/abs/1606.01946

[2] M. Gastpar, B. Rimoldi, and M. Vetterli, "To code, or not to code: Lossy source-channel communication revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1147–1158, 2003.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 2012.

[4] R. Fox and N. Tishby, "Minimum-information LQG control — Part II: Retentive controllers," in *Proceedings of the 55th IEEE Conference on Decision and Control (CDC)*, 2016. [Online]. Available: https://arxiv.org/abs/1606.01947

[5] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby, "Multivariate information bottleneck," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 2001, pp. 152–161.

[6] J. L. Massey, "Causality, feedback and directed information," in *Proceedings of the International Symposium on Information Theory and Its Applications*, 1990, pp. 303–305.

[7] N. Tishby and D. Polani, "Information theory of decisions and actions," in *Perception-Action Cycle*, ser. Cognitive and Neural Systems. Springer, 2011, pp. 601–636.