Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

General duality between optimal control and estimation

Emanuel Todorov

Abstract— **Optimal control and estimation are dual in the LQG setting, as Kalman discovered, however this duality has proven difficult to extend beyond LQG. Here we obtain a more natural form of LQG duality by replacing the Kalman-Bucy filter with the information filter. We then generalize this result to non-linear stochastic systems, discrete stochastic systems, and deterministic systems. All forms of duality are established by relating exponentiated costs to probabilities. Unlike the LQG setting where control and estimation are in one-to-one correspondence, in the general case control turns out to be a larger problem class than estimation and only a sub-class of control problems have estimation duals. These are problems where the Bellman equation is intrinsically linear. Apart from their theoretical significance, our results make it possible to apply estimation algorithms to control problems and vice versa.**

I. Introduction

The best-known example of estimation-control duality is the duality between the Kalman filter and the linear-quadratic regulator. This result was first described in the seminal paper introducing the Kalman filter [6], however it has proven difficult to generalize beyond the linear-quadratic-Gaussian (LQG) setting. Here we develop several such generalizations. The paper is organized as follows. In Section II we show that Kalman's duality is an artifact of the LQG setting, and obtain a new duality which involves the information filter rather than the Kalman filter. In Section III we generalize our new duality to non-linear dynamics and measurements and non-quadratic costs. In Section IV we give a further generalization to discrete dynamics – which can be reduced to the continuous case in section III by assuming Gaussian noise and taking a certain limit. In Section V we develop similar results for deterministic optimal control problems. In Section VI we provide closing remarks and clarify the relations to prior work.

A. Preview of results

Before delving into details we outline the main ideas. All forms of duality we develop here are based on the following relationship between probabilities and costs:

$$r(x,t) \propto \exp(-v(x,t)) \tag{1}$$

$v(x,t)$ is the optimal cost-to-go, i.e. the cost expected to accumulate if we initialize the system in state x at time t and control it optimally until a final time t_f. For discrete systems the optimal cost-to-go satisfies the Bellman equation

$$v(x,t) = \min_u \left\{ \ell(x,u,t) + \sum_{x'} p(x'|x,u)\, v(x',t+1) \right\} \tag{2}$$

where ℓ is the cost rate and $p(x'|x,u)$ is the probability of a transition from state x to state x' under control u.

$r(x,t) = p\left(y_t \cdots y_{t_f} | x_t = x\right)$ is the backward filtering density, i.e. the probability of the future measurements given the current state. For a Markov system it satisfies

$$r(x,t) = p(y_t|x) \sum_{x'} \overline{p}(x'|x)\, r(x',t+1) \tag{3}$$

where $\overline{p}$ is the transition probability without controls (i.e. the passive dynamics) and $p(y_t|x)$ is the emission probability.

The control problem is more general than the estimation problem because of the presence of u in (2). Thus, in order to establish duality, the control problem has to be restricted. The necessary restriction will turn out to be

$$\ell(x,u,t) = -\log p(y_t|x) + \mathrm{KL}\left(p(\cdot|x,u) \,||\, \overline{p}(\cdot|x)\right) \tag{4}$$

The first term is a state cost encouraging the controller to visit more likely states. The second term (Kullback-Liebler divergence between the controlled and passive dynamics) is a control cost encouraging the controller to let the system evolve according to its passive dynamics. With ℓ as in (4) and some additional assumptions, the minimization over u in (2) can be carried out in closed form and, after exponentiation, (2) can be reduced to (3). This is developed in section IV.

The continuous-time results in sections II and III are in some sense special cases, although they will be derived in very different ways and the relation to the discrete case will not become obvious until later. For both linear and non-linear systems subject to Gaussian noise, the KL divergence in (4) will turn out to be identical to a quadratic control cost.

The backward filtering density r in the continuous case is somewhat complicated (see [9]) because a proper density over the space of continuous-time observation sequences is hard to define. Nevertheless r has an intuitive property identical to the discrete case. Let $f(x,t) = p(x_t = x|y_1 \cdots y_{t-1})$ denote the forward filtering density. The product of the forward and backward filtering densities is proportional to the full posterior given all the measurements:

$$p\left(x_t = x | y_1 \cdots y_{t_f}\right) \propto f(x,t)\, r(x,t) \tag{5}$$

The same relationship holds in continuous time.

Emanuel Todorov is with the Department of Cognitive Science, University of California San Diego, todorov@cogsci.ucsd.edu

This work was supported by the US National Science Foundation.
Thanks to Yuval Tassa for his comments on the manuscript.

II. Duality for linear systems

A. Kalman's duality

First we recall Kalman's duality between optimal control and estimation for continuous-time LQG systems. The stochastic dynamics for the control problem are

$$d\mathbf{x} = (A\mathbf{x} + B\mathbf{u})\,dt + Cd\boldsymbol{\omega} \tag{6}$$

The cost accumulates at rate

$$\ell(\mathbf{x}, \mathbf{u}) = \frac{1}{2}\mathbf{x}^{\mathsf{T}} Q\mathbf{x} + \frac{1}{2}\mathbf{u}^{\mathsf{T}} R\mathbf{u} \tag{7}$$

until final time t_f. For simplicity we will assume throughout the paper that there is no final cost, although a final cost can be added and the results still hold. The optimal cost-to-go $v(\mathbf{x}, t)$ for this problem is known to be quadratic. Its Hessian $V(t)$ satisfies the continuous-time Riccati equation

$$-\dot{V} = Q + A^{\mathsf{T}}V + VA - VBR^{-1}B^{\mathsf{T}}V \tag{8}$$

The stochastic dynamics for the dual estimation problem are the same as (6) but with $\mathbf{u} = 0$, namely

$$d\mathbf{x} = A\mathbf{x}dt + Cd\boldsymbol{\omega} \tag{9}$$

The state is now hidden and we have measurement

$$d\mathbf{y} = H\mathbf{x}dt + Dd\boldsymbol{\nu} \tag{10}$$

In discrete time we can write $\mathbf{y}(t) = H\mathbf{x}(t) +$ "noise" because the noise is finite, but here we have the problem that $\dot{\boldsymbol{\nu}}$ is infinite. Therefore the $\mathbf{y}(t)$ defined in (10) is the time-integral of the instantaneous measurements.

Suppose the prior $f(\mathbf{x}, 0)$ over the initial state is Gaussian. Then the forward filtering density $f(\mathbf{x}, t)$ remains Gaussian for all t. Its covariance matrix $\Sigma(t)$ satisfies the continuous-time Riccati equation

$$\dot{\Sigma} = CC^{\mathsf{T}} + A\Sigma + \Sigma A^{\mathsf{T}} - \Sigma H^{\mathsf{T}}\left(DD^{\mathsf{T}}\right)^{-1} H\Sigma \tag{11}$$

Comparing the Riccati equations for the linear-quadratic regulator (8) and the Kalman-Bucy filter (11), we obtain Kalman's duality in continuous time:

linear-quadratic regulator	**Kalman-Bucy filter**
V	Σ
A	A^{T}
B	H^{T}
R	DD^{T}
Q	CC^{T}
t	$t_f - t$

(12)

B. Why Kalman's duality does not generalize

Kalman's duality has been known for half a century and has attracted a lot of attention. If a straightforward generalization to non-LQG settings was possible it would have been discovered long ago. Indeed we will now show that Kalman's duality, although mathematically sound, is an artifact of the LQG setting and needs to be revised before generalizations become possible.

The most obvious problem are the matrix transposes A^{T} and H^{T} in (12). To see the problem consider replacing the linear drift $A\mathbf{x}$ in the controlled dynamics (6) with a general non-linear function $\mathbf{a}(\mathbf{x})$. What is the corresponding change in the estimation dynamics (9)? More precisely, what is the "dual" function $\mathbf{a}^*(\mathbf{x})$ such that $\mathbf{a}(\mathbf{x})$ and $\mathbf{a}^*(\mathbf{x})$ are related in the same way that $A\mathbf{x}$ and $A^{\mathsf{T}}\mathbf{x}$ are related? This question does not appear to have a sensible answer. Generalizing the relationship between B and H^{T} is equally problematic.

The less obvious but perhaps deeper problem is the correspondence between V and Σ. This correspondence may seem related to the exponential transformation (1) between costs and densities, however it is the wrong relationship. If (1) were to hold, the Hessian of $-\log f$ should coincide with V. For Gaussian f the Hessian of $-\log f$ is Σ^{-1}. Thus the general exponential transformation (1) implies a correspondence between V and Σ^{-1}, while in (12) we see a correspondence between V and Σ.

This analysis not only reveals why Kalman's duality does not generalize but also suggests how it should be revised. We need an estimator which propagates Σ^{-1} rather than Σ, i.e. we need an information filter.

C. New duality based on the information filter

The information filter is usually derived in discrete time and its relationship to the linear-quadratic regulator is not obvious. However it can also be derived in continuous time, revealing a new form of estimation-control duality. We use the fact that, if $\Sigma(t)$ is a symmetric positive definite matrix, the time-derivative of its inverse is

$$\frac{d}{dt}\left(\Sigma(t)^{-1}\right) = -\Sigma(t)^{-1}\dot{\Sigma}(t)\Sigma(t)^{-1} \tag{13}$$

Define the inverse covariance matrix $S(t) = \Sigma(t)^{-1}$ and apply (13) to obtain

$$\dot{S}(t) = -S(t)\dot{\Sigma}(t)S(t) \tag{14}$$

Next express $\dot{\Sigma}$ in terms of S by replacing Σ with S^{-1} in the Riccati equation (11). The result is

$$\dot{\Sigma} = CC^{\mathsf{T}} + AS^{-1} + S^{-1}A^{\mathsf{T}} - S^{-1}H^{\mathsf{T}}\left(DD^{\mathsf{T}}\right)^{-1} HS^{-1} \tag{15}$$

Substituting (15) into (14), carrying out the multiplications by S and noting that a number of S and S^{-1} terms cancel, we obtain a continuous-time Riccati equation for S:

$$\dot{S} = H^{\mathsf{T}}\left(DD^{\mathsf{T}}\right)^{-1} H - A^{\mathsf{T}}S - SA - SCC^{\mathsf{T}}S \tag{16}$$

Comparison of (8) and (16) yields our new duality for continuous-time LQG problems:

linear-quadratic regulator	**information filter**
V	Σ^{-1}
A	$-A$
$BR^{-1}B^{\mathsf{T}}$	CC^{T}
Q	$H^{\mathsf{T}}\left(DD^{\mathsf{T}}\right)^{-1} H$
t	$t_f - t$

(17)

As expected we now have a correspondence between V and Σ^{-1}, which is a special case of the exponential transformation (1). The problematic matrix transpose A^{T} from (12) has been replaced with $-A$ which implies a time reversal. This cancels the second time reversal resulting from the different signs of the left hand sides of (16) and (8). Another notable difference is the rearrangement of terms which leads to a very different correspondence between estimation and control. In Kalman's duality the control (B, R) corresponds to the measurement (H, D) while the state cost (Q) corresponds to the dynamics noise (C). Here the control corresponds to the dynamics noise while the state cost corresponds to the measurement, in agreement with (4).

Before proceeding with generalizations we pause to make our new duality more precise. So far all we did was match terms in Riccati equations. However we can now do better: we can identify control and estimation problems whose optimal cost-to-go $v(\mathbf{x}, t)$ and backward filtering density $r(\mathbf{x}, t)$ are related according to (1). The result is as follows:

Theorem 1. *Let $v(\mathbf{x}, t)$ denote the optimal cost-to-go for control problem (6, 7). Let $r(\mathbf{x}, t)$ denote the backward filtering density for estimation problem (9, 10). If all measurements are 0 and*

$$\begin{aligned} BR^{-1}B^{\mathsf{T}} &= CC^{\mathsf{T}} \\ Q &= H^{\mathsf{T}}\left(DD^{\mathsf{T}}\right)^{-1}H \end{aligned} \tag{18}$$

then there exists a positive scalar $c(t)$ such that

$$r(\mathbf{x}, t) = c(t)\exp\left(-v(\mathbf{x}, t)\right) \tag{19}$$

Key to this result is the relationship $V(t) = \Sigma(t)^{-1}$, which follows from the equivalence of the Riccati equations (16) and (8) under (17), and in turn implies (19). Theorem 1 is a special case of Theorem 2 which we prove below. The case of non-zero measurements will also be handled later.

III. Duality for non-linear systems

A. Generalizing the linear results

We now analyze our new duality and infer the form of the non-linear estimation and control problems which are likely to be dual to each other. The correspondence between A and $-A$ implies a time reversal. The last row of (17) is another time reversal, so we can expect the two to cancel. Therefore both the estimation and control problems could have non-linear drift $\mathbf{a}(\mathbf{x})$ instead of $A\mathbf{x}$. The term $BR^{-1}B^{\mathsf{T}}$ suggests that the matrices B and R should be preserved in the generalized problem, that is, we should still have control-affine dynamics and control-quadratic cost. The only possible generalization here is to make B, R, C dependent on $\mathbf{x}$:

$$B(\mathbf{x})R(\mathbf{x})^{-1}B(\mathbf{x})^{\mathsf{T}} = C(\mathbf{x})C(\mathbf{x})^{\mathsf{T}} \tag{20}$$

Next consider the correspondence between Q and $H^{\mathsf{T}}H$. For simplicity we assume that D is the identity matrix although the general case can also be handled. The above correspondence implies correspondence between the quadratic forms $\mathbf{x}^{\mathsf{T}}Q\mathbf{x}$ and $\mathbf{x}^{\mathsf{T}}H^{\mathsf{T}}H\mathbf{x}$. The former equals twice the state-dependent cost, which can be replaced with a general non-quadratic function $q(\mathbf{x})$. The latter involves the linear observation $H\mathbf{x}$, which can be replaced with a general non-linear function $\mathbf{h}(\mathbf{x})$. Then (17) implies

$$q(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{h}(\mathbf{x})\right\|^2 \tag{21}$$

In summary, our analysis suggests controlled dynamics

$$d\mathbf{x} = \left(\mathbf{a}(\mathbf{x}) + B(\mathbf{x})\mathbf{u}\right)dt + C(\mathbf{x})\,d\boldsymbol{\omega} \tag{22}$$

and cost rate

$$\ell(\mathbf{x}, \mathbf{u}) = q(\mathbf{x}) + \frac{1}{2}\mathbf{u}^{\mathsf{T}}R(\mathbf{x})\mathbf{u} \tag{23}$$

For the estimation problem we have dynamics

$$d\mathbf{x} = \mathbf{a}(\mathbf{x})\,dt + C(\mathbf{x})\,d\boldsymbol{\omega} \tag{24}$$

and measurements

$$d\mathbf{y} = \mathbf{h}(\mathbf{x})\,dt + d\boldsymbol{\nu} \tag{25}$$

The generalized duality can now be stated as follows:

Theorem 2. *Let $v(\mathbf{x}, t)$ denote the optimal cost-to-go for control problem (22, 23). Let $r(\mathbf{x}, t)$ denote the backward filtering density for estimation problem (24, 25). If all measurements are 0 and conditions (20, 21) hold, then there exists a positive scalar $c(t)$ such that*

$$r(\mathbf{x}, t) = c(t)\exp\left(-v(\mathbf{x}, t)\right) \tag{26}$$

To prove this theorem we will derive 2nd-order linear PDEs for r and $\exp(-v)$ and show that they are identical. Each PDE is derived in a separate subsection below.

B. Linear Hamilton-Jacobi-Bellman equation

The optimal cost-to-go is known to satisfy the Hamilton-Jacobi-Bellman (HJB) equation. For optimal control problems of the form (22, 23) the HJB equation is

$$\begin{aligned} -v_t = \min_{\mathbf{u}}\Big\{ & q + \frac{1}{2}\mathbf{u}^{\mathsf{T}}R\mathbf{u} + \left(\mathbf{a} + B\mathbf{u}\right)^{\mathsf{T}}v_{\mathbf{x}} \\ & + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}}v_{\mathbf{xx}}\right)\Big\} \end{aligned} \tag{27}$$

The dependence on $(\mathbf{x}, t)$ is suppressed for clarity and subscripts are used to denote partial derivatives. The minimization over $\mathbf{u}$ can be performed in closed form to yield the optimal feedback control law

$$\boldsymbol{\pi}(\mathbf{x}, t) = -R(\mathbf{x})^{-1}B(\mathbf{x})^{\mathsf{T}}v_{\mathbf{x}}(\mathbf{x}, t) \tag{28}$$

Substituting in (27) and dropping the min operator, we obtain the minimized HJB equation

$$-v_t = q + \mathbf{a}^{\mathsf{T}}v_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}}v_{\mathbf{xx}}\right) - \frac{1}{2}v_{\mathbf{x}}^{\mathsf{T}}BR^{-1}B^{\mathsf{T}}v_{\mathbf{x}} \tag{29}$$

Recall that we seek a PDE for $\exp(-v)$ rather than v. To this end we define the exponentially-transformed optimal cost-to-go function

$$z(\mathbf{x}, t) = \exp\left(-v(\mathbf{x}, t)\right) \tag{30}$$

The derivatives of v can be expressed in terms of the derivatives of z:

$$v_t = -\frac{z_t}{z}, \quad v_{\mathbf{x}} = -\frac{z_{\mathbf{x}}}{z}, \quad v_{\mathbf{xx}} = -\frac{z_{\mathbf{xx}}}{z} + \frac{z_{\mathbf{x}} z_{\mathbf{x}}^{\mathsf{T}}}{z^2} \tag{31}$$

Substituting in (29), multiplying by $-z$, and using the properties of the trace operator yields

$$\begin{aligned} -z_t &= -qz + \mathbf{a}^{\mathsf{T}} z_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}} z_{\mathbf{xx}}\right) \\ &\quad + \frac{1}{2z} z_{\mathbf{x}}^{\mathsf{T}} CC^{\mathsf{T}} z_{\mathbf{x}} - \frac{1}{2z} z_{\mathbf{x}}^{\mathsf{T}} BR^{-1} B^{\mathsf{T}} z_{\mathbf{x}} \end{aligned} \tag{32}$$

The last two terms which are quadratic in $z_{\mathbf{x}}$ cancel because of (20). Thus $z(\mathbf{x}, t)$ satisfies the PDE

$$-z_t = -qz + \mathbf{a}^{\mathsf{T}} z_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}} z_{\mathbf{xx}}\right) \tag{33}$$

This is a 2nd-order linear PDE. Note that condition (20), which came from our analysis of duality, was key to cancelling the nonlinear terms and making (33) linear.

C. Backward Zakai equation

The backward filtering density for estimation problems in the form (24, 25) is known to satisfy the backward Zakai equation. More precisely, there exists a positive function $n(\mathbf{x}, t)$ proportional to $r(\mathbf{x}, t)$ which satisfies

$$-dn = \left(\mathbf{a}^{\mathsf{T}} n_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}} n_{\mathbf{xx}}\right)\right) dt + n\mathbf{h}^{\mathsf{T}} d\mathbf{y} \tag{34}$$

The first term on the right corresponds to the backward Kolmogorov equation – which describes how probability densities evolve over time in the absence of measurements. The second term takes into account the measurements.

Equation (34) is a stochastic PDE. In order to transform it into a regular PDE (i.e. put it in a so-called robust form) we follow the approach of [9]. That paper allows the function $\mathbf{h}(\mathbf{x}, t)$ to depend on time and defines

$$\begin{aligned} \epsilon(\mathbf{x}, t) &= \exp\left(\int_0^t \mathbf{h}(\mathbf{x}, s)^{\mathsf{T}} d\mathbf{y}(s) \right. \\ &\quad \left. - \frac{1}{2}\int_0^t \|\mathbf{h}(\mathbf{x}, s)\|^2 ds\right) \end{aligned} \tag{35}$$

It is then shown [9] that

$$-\frac{\partial(n\epsilon)}{\partial t} = \left(\mathbf{a}^{\mathsf{T}} n_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}} n_{\mathbf{xx}}\right)\right)\epsilon \tag{36}$$

In our case $\mathbf{h}$ does not depend on time so ϵ simplifies to

$$\epsilon(\mathbf{x}, t) = \exp\left(\mathbf{h}(\mathbf{x})^{\mathsf{T}}(\mathbf{y}(t) - \mathbf{y}(0)) - \frac{t}{2}\|\mathbf{h}(\mathbf{x})\|^2\right) \tag{37}$$

Now suppose the measurements are $\mathbf{y}(t) = 0$ for all t. This results in further simplification:

$$\begin{aligned} \epsilon(\mathbf{x}, t) &= \exp\left(-\frac{t}{2}\|\mathbf{h}(\mathbf{x})\|^2\right) \\ \frac{\partial \epsilon}{\partial t} &= -\frac{1}{2}\|\mathbf{h}(\mathbf{x})\|^2 \epsilon \end{aligned} \tag{38}$$

Combining (36) and (38) and dividing by ϵ yields

$$-n_t = -\frac{1}{2}\|\mathbf{h}(\mathbf{x})\|^2 n + \mathbf{a}^{\mathsf{T}} n_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}} n_{\mathbf{xx}}\right) \tag{39}$$

Using the relation (21) between $q(\mathbf{x})$ and $\mathbf{h}(\mathbf{x})$ we obtain

$$-n_t = -qn + \mathbf{a}^{\mathsf{T}} n_{\mathbf{x}} + \frac{1}{2}\operatorname{tr}\left(CC^{\mathsf{T}} n_{\mathbf{xx}}\right) \tag{40}$$

The latter PDE is identical to (33). This completes the proof of Theorem 2.

We mentioned earlier that our results can be generalized to non-zero measurements. Indeed, if $\mathbf{y}(t)$ is any differentiable function of t, repeating the above derivation yields the following generalization to (21):

$$q(\mathbf{x}, t) = \frac{1}{2}\|\mathbf{h}(\mathbf{x})\|^2 - \mathbf{h}(\mathbf{x})^{\mathsf{T}} \dot{\mathbf{y}}(t) \tag{41}$$

Thus q in general depends on the measurements. This is to be expected: to establish duality as outlined in the introduction we need a cost q penalizing unlikely states, and the likelihood of the states depends on the measurements.

IV. Duality for discrete systems

This section develops an estimation-control duality for a new class of Markov decision problems (MDPs) which we recently introduced [13]. Below we first summarize the relevant properties of these MDPs, and then establish a duality to hidden Markov models (HMMs). We also show how the continuous control problems in the previous section can be obtained from these MDPs by taking a certain limit.

A. Linearly-solvable MDPs

Consider a standard MDP setting where $p(x'|x, u)$ is the probability of a transition from state x to state x' under control u, and $\ell(x, u)$ the cost for being in state x and choosing control u. As stated in the introduction (using slightly different notation), the optimal cost-to-go satisfies the Bellman equation

$$v(x, t) = \min_u \left\{\ell(x, u) + \mathrm{E}_{x' \sim p(\cdot|x,u)}\left[v(x', t+1)\right]\right\} \tag{42}$$

For standard MDPs the Bellman equation requires exhaustive search over the set of admissible controls for each x. In order to avoid this inefficiency, we recently introduced a new class of MDPs where the search is replaced with an analytical solution [13]. The controls in these new MDPs directly specify the transition probabilities:

$$p(x'|x, u(\cdot)) = u(x') \tag{43}$$

Each control $u(\cdot)$ is a collection of non-negative real numbers which sum to 1. We constrain the controls by introducing the notion of passive/uncontrolled dynamics $\overline{p}(x'|x)$ and requiring the controls to be compatible with $\overline{p}$ as follows:

$$\text{if } \ \overline{p}(x'|x) = 0 \ \text{ then we require } \ u(x') = 0 \tag{44}$$

We also constrain the cost function $\ell(x, u)$ to the form

$$\begin{aligned} \ell(x, u(\cdot)) &= q(x) + \mathrm{KL}\left(u(\cdot) \,||\, \overline{p}(\cdot|x)\right) \\ &= q(x) + \mathrm{E}_{x' \sim u(\cdot)}\left[\log \frac{u(x')}{\overline{p}(x'|x)}\right] \end{aligned} \tag{45}$$

$q(x) \geq 0$ can be any scalar function encoding how (un)desirable different states are. The KL divergence plays the role of a control cost and penalizes the difference between the controlled and passive dynamics.

For the above class of MDPs the Bellman equation is

$$\begin{aligned} v(x,t) &= \min_{u(\cdot)} \{q(x) + \mathrm{KL}(u(\cdot)\,||\,\overline{p}(\cdot|x)) \\ &\qquad + \mathrm{E}_{x'\sim u(\cdot)}[v(x',t+1)]\} \\ &= q(x) - \log(\text{normalizer}) \\ &\quad + \min_{u(\cdot)} \mathrm{KL}\left(u(\cdot)\,\middle\|\,\frac{\overline{p}(\cdot|x)\exp(-v(\cdot,t+1))}{\text{normalizer}}\right) \\ &= q(x) - \log \mathrm{E}_{x'\sim\overline{p}(\cdot|x)}[\exp(-v(x',t+1))] \end{aligned} \tag{46}$$

The transformation from the 1st to the 2nd line is straightforward [13]. The minimum of the KL divergence is 0 and is achieved when the two distributions are equal – which yields the 3d line above as well as the optimal control law:

$$u^*_{x,t}(\cdot) \propto \overline{p}(\cdot|x)\exp(-v(\cdot,t+1)) \tag{47}$$

As before we seek an equation for the exponentially-transformed optimal cost-to-go function

$$z(x,t) = \exp(-v(x,t)) \tag{48}$$

Exponentiating (46) and expressing it in terms of z yields

$$z(x,t) = \exp(-q(x))\,\mathrm{E}_{x'\sim\overline{p}(\cdot|x)}[z(x',t+1)] \tag{49}$$

Note that we have not only replaced the exhaustive search over controls with an analytical solution but also transformed the Bellman equation into a linear equation.

B. Duality between HMMs and our MDPs

The transformed Bellman equation (49) has the same from as equation (3) which governs the backward filtering density for HMMs. This suggests a duality between our MDPs and HMMs, as follows. On the control side we have dynamics

$$x_{t+1} \sim u(\cdot|x_t) \tag{50}$$

and cost function

$$\ell(x,u) = q(x) + \mathrm{KL}(u(\cdot|x)\,||\,\overline{p}(\cdot|x)) \tag{51}$$

On the estimation side we have dynamics

$$x_{t+1} \sim \overline{p}(\cdot|x_t) \tag{52}$$

and binary measurements with emission probability

$$p(y_t = 0|x_t = x) = g(x) \tag{53}$$

The duality can now be stated as follows:

Theorem 3. *Let $v(x,t)$ denote the optimal cost-to-go for control problem (50, 51). Let $r(x,t)$ denote the backward filtering density for estimation problem (52, 53). If all measurements are 0 and*

$$q(x) = -\log(g(x)) \tag{54}$$

then there exists a positive scalar $c(t)$ such that

$$r(x,t) = c(t)\exp(-v(x,t)) \tag{55}$$

This theorem follows from the fact that the solutions to the above control and estimation problems satisfy identical equations: (49) and (3) respectively.

C. Relationship to our continuous problems

Here we relate the above MDPs to the continuous control problems (22, 23) from the previous section. This is done in two steps. First we make the state space Euclidean and define a family of continuous-space discrete-time problems indexed by the discrete time step $h > 0$. Then we take a continuous-time limit $\lim_{h\downarrow 0}$.

Let $\overline{p}^{(h)}(\mathbf{x}'|\mathbf{x})$ denote the passive dynamics, that is, the probability of being in state $\mathbf{x}'$ at time h given that the system was initialized in state $\mathbf{x}$ at time 0. Denote the exponentially-transformed optimal cost-to-go for this problem with $z^{(h)}(\mathbf{x},t)$ where t is an integer multiple of h. Computing $z^{(h)}$ is identical to our derivation in the MDP case except that all sums are now replaced with integrals. The linear Bellman equation becomes

$$z^{(h)}(\mathbf{x},t) = \exp(-q(\mathbf{x})\,h)\,\mathrm{E}_{\mathbf{x}'\sim\overline{p}^{(h)}(\cdot|\mathbf{x})}\left[z^{(h)}(\mathbf{x}',t+h)\right] \tag{56}$$

The state cost is now $q(\mathbf{x})\,h$ because the cost accumulates over time period h at rate $q(\mathbf{x})$. Define $z = \lim_{h\downarrow 0} z^{(h)}$. In order to derive a PDE characterizing z, we multiply by $\exp(qh)$, subtract $z^{(h)}$, divide by h and take the limit:

$$\begin{aligned} &\lim_{h\downarrow 0} \frac{\exp(q(\mathbf{x})\,h) - 1}{h} z^{(h)}(\mathbf{x},t) = \\ &\lim_{h\downarrow 0} \frac{\mathrm{E}_{\mathbf{x}'\sim\overline{p}^{(h)}(\cdot|\mathbf{x})}\left[z^{(h)}(\mathbf{x}',t+h) - z^{(h)}(\mathbf{x},t)\right]}{h} \end{aligned} \tag{57}$$

The first limit evaluates to qz. The second limit coincides with the notion of generalized derivative in the theory of stochastic processes and evaluates to $z_t + \mathcal{L}[z]$, where the operator $\mathcal{L}$ is the infinitesimal generator [12] of the stochastic process with transition probability $\overline{p}^{(h)}(\mathbf{x}'|\mathbf{x})$. Thus $-z_t = -qz + \mathcal{L}[z]$. The generator $\mathcal{L}$ of course depends on the passive dynamics. For a diffusion process of the form (24) the generator is known to be

$$\mathcal{L}[z] = \mathbf{a}^{\mathsf{T}} z_{\mathbf{x}} + \frac{1}{2}\,\mathrm{tr}\left(CC^{\mathsf{T}} z_{\mathbf{xx}}\right) \tag{58}$$

Putting these results together we obtain the PDE

$$-z_t = -qz + \mathbf{a}^{\mathsf{T}} z_{\mathbf{x}} + \frac{1}{2}\,\mathrm{tr}\left(CC^{\mathsf{T}} z_{\mathbf{xx}}\right) \tag{59}$$

which is identical to (33) in the previous section. Thus our new MDPs represent a generalization of problem (22, 23).

Recall that in our MDPs the control is a probability distribution over reachable states. When the state space is made continuous the control should become infinite dimensional (i.e. a probability density). But if this is so, how did we recover (22, 23) which involves finite-dimensional control? The answer is that, although in principle the control can be any probability density, the optimal control is a shifted version of $\overline{p}$ and so we can parameterize it with the vector $\mathbf{u}$ in (22, 23). This is because for small h the density $\overline{p}^{(h)}(\cdot|\mathbf{x})$ is sharply peaked and approximately Gaussian, and multiplication by a smooth $\exp(-v(\cdot))$ as in (47) can do

nothing more than shift the mean of that Gaussian. The latter statement holds only to first order in h, but in the continuous-time limit first order in h is all that matters.

The relation we established between our MDPs and problems of the form (22, 23) suggests that KL divergences and quadratic control costs are related. To see why, note that for small h the transition probability densities for both the controlled and the passive dynamics are approximately Gaussian, with covariance $hB(\mathbf{x})B(\mathbf{x})^{\mathsf{T}}$ and means which differ by $hB(\mathbf{x})\mathbf{u}$. Applying the standard formula for KL divergence between Gaussians yields control cost $\frac{h}{2}\|\mathbf{u}\|^2$ per time h, and so the control cost rate is $\frac{1}{2}\|\mathbf{u}\|^2$.

V. Duality for deterministic systems

The duality results presented thus far were obtained by defining pairs of optimal control and estimation problems, deriving equations that characterize $\exp(-v)$ and r, and showing that these equations are identical. This indirect approach was needed because we were interested in filtering densities which are not defined as the solution to an optimization problem (but see [10]). However if we are only interested in the peak of the density – as in maximum *a posteriori* (MAP) estimation – then the estimation problem is formulated in terms of optimization and can be directly converted into an optimal control problem, without having to characterize the solution to either problem. This is the approach we take here. Another important difference here is that (point) estimation will turn out to be dual to deterministic optimal control. First we give results for general non-linear systems and then specialize them to the linear case. The states, controls and measurements in this section are real-valued vectors while the time is discrete.

A. MAP smoothing and deterministic control

Consider a partially observable stochastic system with transition probability function $\overline{p}$ and emission probability function $p_{\mathbf{y}}$ defined as

$$\begin{aligned}\overline{p}(\mathbf{x}'|\mathbf{x}) &= \exp(-k(\mathbf{x}',\mathbf{x})) \\ p_{\mathbf{y}}(\mathbf{y}|\mathbf{x}) &= \exp(-q(\mathbf{y},\mathbf{x}))\end{aligned} \tag{60}$$

k, q are known scalar functions. Suppose we are given a sequence of observations $(\mathbf{y}_1, \cdots \mathbf{y}_{n-1})$ denoted $\mathbf{y}_{1:n-1}$. Our objective is to find the most probable sequence of states $(\mathbf{x}_1, \cdots \mathbf{x}_n)$, that is, the sequence which maximizes the posterior probability

$$p(\mathbf{x}_{1:n}|\mathbf{y}_{1:n-1}) = \frac{p(\mathbf{y}_{1:n-1}|\mathbf{x}_{1:n})\,p(\mathbf{x}_{1:n})}{p(\mathbf{y}_{1:n-1})} \tag{61}$$

The denominator does not affect the maximization so it can be ignored. Assuming an uninformative prior over $\mathbf{x}_1$ and using the Markov property of (60) we have

$$\begin{aligned}&p(\mathbf{y}_{1:n-1}|\mathbf{x}_{1:n})\,p(\mathbf{x}_{1:n}) \\ &= \textstyle\prod_{t=1}^{n-1} p_{\mathbf{y}}(\mathbf{y}_t|\mathbf{x}_t)\,\overline{p}(\mathbf{x}_{t+1}|\mathbf{x}_t) \\ &= \exp\Big(-\textstyle\sum_{t=1}^{n-1} q(\mathbf{y}_t,\mathbf{x}_t) + k(\mathbf{x}_{t+1},\mathbf{x}_t)\Big)\end{aligned} \tag{62}$$

Maximizing the above expression is equivalent to minimizing its negative log, which we denote with J:

$$J(\mathbf{x}_{1:n}) = \sum\nolimits_{t=1}^{n-1} q(\mathbf{y}_t,\mathbf{x}_t) + k(\mathbf{x}_{t+1},\mathbf{x}_t) \tag{63}$$

This is beginning to look like a total cost for an optimal control problem with state cost q and control cost k. However we are still missing an explicit control signal. To remedy that we define the (deterministic) controlled dynamics as

$$\mathbf{x}_{t+1} = \mathbf{a}(\mathbf{x}_t) + \mathbf{u}_t \tag{64}$$

where $\mathbf{a}(\mathbf{x})$ is the expected next state under $\overline{p}$:

$$\mathbf{a}(\mathbf{x}) = \mathrm{E}_{\mathbf{x}'\sim\overline{p}(\cdot|\mathbf{x})}[\mathbf{x}'] \tag{65}$$

The results below actually hold regardless of how we define $\mathbf{a}(\mathbf{x})$, yet the present definition is the most intuitive. The control $\mathbf{u}$ is a perturbation to the passive dynamics $\mathbf{a}(\mathbf{x})$.

The cost for the control problem will be defined as

$$\ell(\mathbf{x},\mathbf{u},t) = q(\mathbf{y}_t,\mathbf{x}) + k(\mathbf{a}(\mathbf{x})+\mathbf{u},\mathbf{x}) \tag{66}$$

The state cost q relates to the emission probability $p_{\mathbf{y}}$ in the same way as it did in Theorem 3. The control cost k is no longer a KL divergence; instead it is the log-likelihood of the perturbation/control. It is now easy to verify that

$$\begin{aligned}J(\mathbf{x}_{1:n}) &= \sum\nolimits_{t=1}^{n-1} \ell(\mathbf{x}_t, \mathbf{x}_{t+1} - \mathbf{a}(\mathbf{x}_t), t) \\ &= \sum\nolimits_{t=1}^{n-1} \ell(\mathbf{x}_t,\mathbf{u}_t,t)\end{aligned} \tag{67}$$

This yields the following result:

Theorem 4. *For any observation sequence, the optimal state trajectory for estimation problem (60) is identical to the optimal state trajectory for control problem (64, 66).*

We assumed an uninformative prior, however the same result holds if an initial state $\mathbf{x}_0$ is given both in the estimation and in the control problem. The only change is the addition of $k(\mathbf{x}_1,\mathbf{x}_0)$ to both sides of (67).

B. The LQG case

Let us now specialize the above results to the LQG setting. The general functions k, q take the specific form

$$\begin{aligned}k(\mathbf{x}',\mathbf{x}) &= \tfrac{1}{2}(\mathbf{x}'-A\mathbf{x})^{\mathsf{T}}\left(CC^{\mathsf{T}}\right)^{-1}(\mathbf{x}'-A\mathbf{x}) + k_0 \\ q(\mathbf{y},\mathbf{x}) &= \tfrac{1}{2}(\mathbf{y}-H\mathbf{x})^{\mathsf{T}}\left(DD^{\mathsf{T}}\right)^{-1}(\mathbf{y}-H\mathbf{x}) + q_0\end{aligned} \tag{68}$$

These functions correspond to a discrete-time estimation problem with linear dynamics

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + C\mathbf{w}_t \tag{69}$$

and linear measurement

$$\mathbf{y}_t = H\mathbf{x}_t + D\mathbf{v}_t \tag{70}$$

where $\mathbf{w}_t, \mathbf{v}_t$ are standard normal random variables. For simplicity we will again assume zero measurements.

The corresponding control problem has dynamics

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B\mathbf{u}_t \tag{71}$$

and cost function

$$\ell(\mathbf{x},\mathbf{u}) = \frac{1}{2}\mathbf{x}^{\mathsf{T}} Q\mathbf{x} + \frac{1}{2}\mathbf{u}^{\mathsf{T}} R\mathbf{u} \tag{72}$$

It is clear that, in order to make the above control problem compatible with the general form (64, 66), the following relations have to hold:

$$\begin{aligned} BR^{-1}B^{\mathsf{T}} &= CC^{\mathsf{T}} \\ Q &= H^{\mathsf{T}}\left(DD^{\mathsf{T}}\right)^{-1} H \end{aligned} \tag{73}$$

These are the same relations we discovered in Section II and generalized in Sections III and IV.

VI. Discussion

Here we obtained a new estimation-control duality in the LQG setting and generalized it to non-linear stochastic systems, discrete stochastic systems and deterministic systems. Some aspects of our work are related to prior developments. The fact that the exponential transformation leads to linear HJB equations is well known [5], [8], [3], [7]. Estimation-control dualities exploiting this fact we studied in [3], however they involved forward filtering instead of backward filtering and as a result were less natural. More recently [10] obtained a form of duality related to Theorem 2, although using a different method. In the context of MAP smoothing our work has similarities with the idea of minimum-energy filters [11]. Researchers in machine learning [1], [14] have used estimation methods to find optimal controls, however these methods operate in the product space of states and controls. In contrast, we perform estimation only in the state space and then use the filtering density to compute optimal controls. Kalman's original duality has been exploited to compute optimal controls for LQG systems with multiple input delays [15]. It will be interesting to see if our general duality can be used to extend these results beyond LQG.

All forms of duality we described here were based on the exponential relationship (1) between probabilities and costs. This fundamental relationship arises in a number of other fields. In statistical physics, (1) is the Gibbs distribution relating the energy $v(x)$ of state x and the probability $r(x)$ of observing the system in state x at thermal equilibrium. In machine learning, (1) relates the model-fitting error $v(x)$ and the likelihood $r(x)$, where x are the model parameters. Indeed most machine learning methods have both error-minimization and likelihood-maximization forms.

While Kalman's original duality suggested that optimal estimation and optimal control are in one-to-one correspondence, our results show that this is generally not the case. The class of stochastic optimal control problems that have estimation duals are those with control-affine dynamics, control-quadratic costs, and dynamics noise satisfying the relationship $BR^{-1}B^{\mathsf{T}} = CC^{\mathsf{T}}$. We saw repeatedly that this relationship was necessary in order to establish duality. The dual estimation problems on the other hand were not constrained – indeed (24, 25) is the general problem of non-linear estimation usually studied in the literature. The fact that a special family of stochastic optimal control problems are dual to a general family of Bayesian estimation problems leads to the conjecture that control problems outside this class may lack estimation duals.

Our results make it possible to develop new algorithms for optimal control by adapting corresponding estimation algorithms. One very popular class of estimation algorithms are particle filters – which represent probability distributions with samples rather than (possibly inaccurate) function approximators. Particle filters do not yet have analogs in the control domain. Our duality makes it possible to obtain such analogs. One complication is that most existing particle filters run forward in time while we need a filter that runs backward in time. Some progress along these lines has been made [4]. Other popular Bayesian inference algorithms include variational approximations and loopy belief propagation in graphical models [2], although these algorithms are usually applied to discrete state spaces.

Finally, our results make it possible to obtain a classic maximum principle for stochastic optimal control problems possessing estimation duals. In particular, we can start with the stochastic control problems in sections III and IV, transform them into dual estimation problems, and transform the latter into deterministic control problems as in section V. Pontryagin's maximum principle can then be applied. These ideas will be developed in future work.

References

[1] H. Attias. Planning by probabilistic inference. *AISTATS*, 2003.

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Spinger, 2007.

[3] W. Fleming and S. Mitter. Optimal control and nonlinear filtering for nondegenerate diffusion processes. *Stochastics*, 8:226–261, 1982.

[4] S. Godsill, A. Doucet, and M. West. Monte carlo smoothing for nonlinear time series. *Journal of the American Statistical Association*, 99:156–168, 2004.

[5] C. Holland. A new energy characterization of the smallest eigenvalue of the schrödinger equation. *Comm Pure Appl Math*, 30:755–765, 1977.

[6] R. Kalman. A new approach to linear filtering and prediction problems. *ASME Transactions journal of basic engineering*, 82(1):35–45, 1960.

[7] H. Kappen. Linear theory for control of nonlinear stochastic systems. *Physical Review Letters*, 95, 2005.

[8] I. Karatzas. On a stochastic representation for the principal eigenvalue of a second-order differential equation. *Stochastics*, 3:305–321, 1980.

[9] V. Krishnamurthy and R. Elliott. Robust continuous-time smoothers without two-sided stochastic integrals. *IEEE Transactions on Automatic Control*, 47:1824–1841, 2002.

[10] S. Mitter and N. Newton. A variational approach to nonlinear estimation. *SIAM J Control Opt*, 42:1813–1833, 2003.

[11] R. Mortensen. Maximum-likelihood recursive nonlinear filtering. *J Optimization Theory and Applications*, 2:386–394, 1968.

[12] B. Oksendal. *Stochastic Differential Equations (4th Ed)*. Springer-Verlag, Berlin, 1995.

[13] E. Todorov. Linearly-solvable markov decision problems. *Advances in Neural Information Processing Systems*, 2006.

[14] M. Toussaint and A. Storkey. Probabilistic inference for solving discrete and continuous state markov decision processes. *International Conference on Machine Learning*, 23, 2006.

[15] H. Zhang, G. Duan, and L. Xie. Linear quadratic regulation for linear time-varying systems with multiple input delays. *Automatica*, 42:1465–1476, 2006.