Gradient Flows in Uncertainty Propagation and Filtering of Linear Gaussian Systems

Abhishek Halder, and Tryphon T. Georgiou
University of California, Irvine

***Abstract*— The purpose of this work is mostly expository and aims to elucidate the Jordan-Kinderlehrer-Otto (JKO) scheme for uncertainty propagation, and a variant, the Laugesen-Mehta-Meyn-Raginsky (LMMR) scheme for filtering. We point out that these variational schemes can be understood as proximal operators in the space of density functions, realizing gradient flows. These schemes hold the promise of leading to efficient ways for solving the Fokker-Planck equation as well as the equations of non-linear filtering. Our aim in this paper is to develop in detail the underlying ideas in the setting of linear stochastic systems with Gaussian noise and recover known results.**

I. Introduction

Consider the gradient flow $\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = -\nabla\psi(\boldsymbol{x})$ in $\mathbb{R}^n$, where ∇ is the gradient (w.r.t. the Euclidean metric) of a function $\psi(\boldsymbol{x})$, and consider the discretization

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - h\nabla\psi(\boldsymbol{x}_{k-1}), \text{ for } k \in \mathbb{N}.$$

As is well known in finite-dimensional optimization,

$$\begin{aligned}\boldsymbol{x}_k &= \underset{\boldsymbol{x}}{\operatorname{argmin}}\{\frac{1}{2}\|\boldsymbol{x} - (\boldsymbol{x}_{k-1} - h\nabla\psi(\boldsymbol{x}_{k-1}))\|^2\} \\ &= \underset{\boldsymbol{x}}{\operatorname{argmin}}\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_{k-1}\|^2 + h\psi(\boldsymbol{x}) + o(h). \qquad (1)\end{aligned}$$

By recursively evaluating the proximal operator

$$\begin{aligned}\boldsymbol{x}_k &= \operatorname{prox}_{h\psi}^{\|\cdot\|}(\boldsymbol{x}_{k-1}) \\ &= \underset{\boldsymbol{x}}{\operatorname{argmin}}\{\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_{k-1}\|^2 + h\psi(\boldsymbol{x})\},\end{aligned}$$

the solution, which depends on the choice of the step size h, satisfies $\boldsymbol{x}_k(h) \to \boldsymbol{x}(t = kh)$, as $h \to 0$.

The Jordan-Kinderlehrer-Otto (JKO) scheme, introduced in [1], is a similar recursion in the infinite-dimensional space of density functions with respect to the Wasserstein geometry [2], namely,

$$\varrho_k(\boldsymbol{x}, h) = \underset{\varrho}{\operatorname{argmin}}\ \frac{1}{2}W_2^2(\varrho, \varrho_{k-1}) + h\mathcal{S}(\varrho), \quad k \in \mathbb{N}, \quad (2)$$

where $W_2(\cdot,\cdot)$ denotes the Wasserstein-2 distance between two (probability) density functions,

$$\mathcal{S}(\varrho) := \int_{\mathbb{R}^n} \varrho(\boldsymbol{x}) \log(\varrho(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \quad (3)$$

is the negative differential entropy functional, and $\mathrm{d}\boldsymbol{x}$ is the volume element. I.e., (2) can be viewed as the proximal operation $\operatorname{prox}_{h\mathcal{S}}^{W_2}(\varrho_{k-1})$. The main result in [1] was to show that the minimizer of (2) approximates the solution $\rho(\boldsymbol{x}, t)$ of the heat equation

$$\frac{\partial\rho(\boldsymbol{x}, t)}{\partial t} = \Delta\rho(\boldsymbol{x}, t), \text{ with } \rho(\boldsymbol{x}, 0) = \rho_0(\boldsymbol{x}),$$

in the sense that $\varrho_k(\boldsymbol{x}, h) \to \rho(\boldsymbol{x}, t = kh)$, as $h \downarrow 0$. Thus, (2) establishes the remarkable result that *the heat equation is the gradient descent flow of the (negative) entropy integral with respect to the Wasserstein metric*.

An analogous JKO-like scheme was introduced recently in Laugesen *et al.* [6] for the measurement update-step in continuous-time filtering. More specifically, we consider the general system of stochastic differential equations (SDE's)

$$\begin{aligned}\mathrm{d}\boldsymbol{x}(t) &= -\nabla U(\boldsymbol{x})\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}\boldsymbol{w}(t), \qquad (4a) \\ \mathrm{d}\boldsymbol{z}(t) &= \boldsymbol{c}(\boldsymbol{x}(t), t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{v}(t), \qquad (4b)\end{aligned}$$

where $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{z} \in \mathbb{R}^m, \beta > 0$, $U(\cdot)$ is a potential, the process and measurement noise processes $\boldsymbol{w}(t)$ and $\boldsymbol{v}(t)$ are Wiener and satisfy $\mathbb{E}\left[\mathrm{d}w_i\mathrm{d}w_j\right] = \boldsymbol{Q}_{ij}\mathrm{d}t\ \forall\ i, j = 1, \ldots, n$ and $\mathbb{E}\left[\mathrm{d}v_i\mathrm{d}v_j\right] = \boldsymbol{R}_{ij}\mathrm{d}t\ \forall\ i, j = 1, \ldots, m$, with $\boldsymbol{Q}, \boldsymbol{R} \succ \boldsymbol{0}$, respectively. Then $\boldsymbol{x}(t)$ and $\boldsymbol{z}(t)$ represent state and sensor measurements at time t. Further, as usual, $\boldsymbol{v}(t)$ is assumed to be independent of $\boldsymbol{w}(t)$ and independent of the initial state $\boldsymbol{x}(0)$. Given the history of noise corrupted sensor data up to time t, the filtering problem requires computing the posterior probability distribution that obeys the Kushner-Stratonovich stochastic PDE [17]–[19].

For the special case of trivial state dynamics, i.e., $\mathrm{d}\boldsymbol{x} = 0$, and $\boldsymbol{R}$ the identity, Laugesen *et al.* [6] introduced

$$\varrho_k^+(\boldsymbol{x}, h) = \underset{\varrho\in\mathcal{D}_2}{\operatorname{arginf}}\{D_{\mathrm{KL}}\left(\varrho\|\varrho_k^-\right) + h\Phi(\varrho)\},\ k \in \mathbb{N}, \quad (5)$$

with

$$\Phi(\varrho) := \frac{1}{2}\mathbb{E}_\varrho\{(\boldsymbol{y}_k - \boldsymbol{c}(\boldsymbol{x}))^\top \boldsymbol{R}^{-1}(\boldsymbol{y}_k - \boldsymbol{c}(\boldsymbol{x}))\}, \quad (6)$$

where $\boldsymbol{y}_k$ is the noisy measurement in discrete-time defined via $\boldsymbol{y}_k := \frac{1}{h}\Delta\boldsymbol{z}_k$, $\Delta\boldsymbol{z}_k := \boldsymbol{z}_k - \boldsymbol{z}_{k-1}$, and $\{\boldsymbol{z}_{k-1}\}_{k\in\mathbb{N}}$ the sequence of samples of $\boldsymbol{z}(t)$ at $\{t_{k-1}\}_{k\in\mathbb{N}}$ for $t_{k-1} := (k-1)h$. Laugesen *et al.* [6] proved that the *LMMR equation* (5) approximates the solution of

$$\begin{aligned}\mathrm{d}\rho^+(\boldsymbol{x}(t), t) = \Big[&\big(\boldsymbol{c}(\boldsymbol{x}(t), t) - \mathbb{E}_{\rho^+}\{\boldsymbol{c}(\boldsymbol{x}(t), t)\}\big)^\top \boldsymbol{R}^{-1} \\ &\big(\mathrm{d}\boldsymbol{z}(t) - \mathbb{E}_{\rho^+}\{\boldsymbol{c}(\boldsymbol{x}(t), t)\}\mathrm{d}t\big)\Big]\ \rho^+(\boldsymbol{x}(t), t), \quad (7)\end{aligned}$$

i.e., of the Kushner-Stratonovich PDE corresponding to $\mathrm{d}\boldsymbol{x} = 0$, in the sense that $\varrho_k^+(\boldsymbol{x}, h) \rightharpoonup \rho^+(\boldsymbol{x}(t), t)$ over $t \in [(k-1)h, kh)$, as $h \downarrow 0$. Thus, they showed that in this special case, *the Kushner-Stratonovich PDE is the gradient descent of functional* $\Phi(\cdot)$ *with respect to* $D_{\rm KL}$*, i.e., computed by* $\mathrm{prox}_{h\Phi}^{D_{\rm KL}}(\varrho_k^-)$.

The purpose of the present paper is to develop this circle of ideas, namely, that *both uncertainty propagation and filtering can be viewed as gradient flows* in the special case of linear stochastic systems with Gaussian noise. In fact, we consider the general case of the linear stochastic system

$$\mathrm{d}\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{x}(t)\,\mathrm{d}t\, + \boldsymbol{B}\,\mathrm{d}\boldsymbol{w}(t), \tag{8}$$

where $\boldsymbol{w}(t)$ is a Wiener process as before, though possibly not of the same dimension as $\boldsymbol{x}$. We suppose that the uncertain initial condition $\boldsymbol{x}(0)$ has a known Gaussian PDF, the matrix $\boldsymbol{A}$ is Hurwitz, and that the diffusion matrix $\boldsymbol{B}$ is such that $(\boldsymbol{A},\boldsymbol{B})$ is a controllable pair. For this, we recover the well-known propagation equations for the mean and covariance of the state $\boldsymbol{x}(t)$ out of the JKO-scheme via a two-step optimization. The applicability of the JKO-scheme to (8) *is not immediately obvious* since the development in [1] requires the state dynamics to be in the canonical form (4a) with the drift being a gradient and the diffusion coefficient being a positive scalar. We further show that this two-step optimization procedure that we introduce, can be used to derive the Kalman-Bucy filter from a generalized version of the LMMR equation (5). We remark that variational schemes for estimator/observer design based on gradient flows can also be seen as regularized dynamic inversion in the spirit of [21].

Notation

Throughout we use bold-faced upper-case letters for matrices, and bold-faced lower case letters for vectors. The notation $\boldsymbol{I}$ stands for identity matrix of appropriate dimension, we use $\mathrm{tr}(\cdot)$ and $\det(\cdot)$ to respectively denote the trace and determinant of a matrix, and the symbols ∇ and $\triangle$ denote the gradient and Laplacian operators, respectively. We denote the space of probability density functions (PDFs) on $\mathbb{R}^n$ by $\mathscr{D} := \{\rho : \rho \geq 0, \int_{\mathbb{R}^n}\rho = 1\}$, by $\mathscr{D}_2 := \{\rho \in \mathscr{D} \mid \int_{\mathbb{R}^n}\boldsymbol{x}^\top\boldsymbol{x}\,\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x} < \infty\}$ the space of PDFs with finite second moments, by $\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}$ denote the space of PDFs which share the same mean vector $\boldsymbol{\mu}$ and same covariance matrix $\boldsymbol{P} := \int_{\mathbb{R}^n}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. Likewise, let $\mathscr{D}_{\boldsymbol{\mu},\tau}$ denote the space of PDFs which have the same mean $\boldsymbol{\mu}$ and same trace of covariance $\tau := \mathrm{tr}(\boldsymbol{P}) > 0$. Clearly, $\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}} \subset \mathscr{D}_{\boldsymbol{\mu},\tau} \subset \mathscr{D}_2 \subset \mathscr{D}$. We use the symbol $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{P})$ to denote a multivariate Gaussian PDF with mean $\boldsymbol{\mu}$, and covariance $\boldsymbol{P}$. The notation $\boldsymbol{x}\sim\rho$ means that the random vector $\boldsymbol{x}$ has PDF ρ; and $\mathbb{E}\{\cdot\}$ denotes the expectation operator while, when the probability density is to be specified, $\mathbb{E}_\rho\{\cdot\} := \int_{\mathbb{R}^n}(\cdot)\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

II. JKO Scheme in General

We now discuss in some detail the JKO scheme for the case of the diffusion process in (4a), and the corresponding Fokker-Planck equation [4]

$$\frac{\partial\rho}{\partial t} = \nabla\cdot(\nabla U(\boldsymbol{x})\rho) + \beta^{-1}\triangle\rho, \quad \rho(\boldsymbol{x},0) = \rho_0(\boldsymbol{x}). \tag{9}$$

To this end we first introduce the Wasserstein metric, the free energy, and the Kullback-Leibler divergence.

The **Wasserstein-2 distance** $W_2(\rho_1,\rho_2)$ between a pair of PDFs $\rho_1(\boldsymbol{x}),\rho_2(\boldsymbol{y}) \in \mathscr{D}$ (or, even between probability measures, in general), supported on $\mathcal{X},\mathcal{Y} \subseteq \mathbb{R}^n$, is

$$W_2(\rho_1,\rho_2) := \left(\inf_{\mathrm{d}\sigma\in\Pi(\rho_1,\rho_2)}\int_{\mathcal{X}\times\mathcal{Y}}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\,\mathrm{d}\sigma(\boldsymbol{x},\boldsymbol{y})\right)^{\frac{1}{2}} \tag{10}$$

where $\Pi(\rho_1,\rho_2)$ is a probability measure on the product space $\mathcal{X}\times\mathcal{Y}$ having finite second moments and marginals ρ_1,ρ_2, respectively. It is well known that $W_2 : \mathscr{D}\times\mathscr{D} \mapsto [0,\infty)$ is a metric [2, p. 208]. Further, its square $W_2^2(\rho_1,\rho_2)$ represents the smallest amount of "work" needed to "morph" ρ_1 into ρ_2 [3]. The infimum is achieved over a space of measures, and under mild assumptions, the minimizing $\mathrm{d}\sigma$ has support on the graph of the optimal "transportation map" $T : \mathcal{X} \mapsto \mathcal{Y}$ that pushes ρ_1 to ρ_2. Alternatively, one may view the optimization problem in (10) as seeking the joint distribution of two random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, distributed according to ρ_1 and ρ_2 respectively, that minimizes the variance $\mathbb{E}\left\{\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right\}$.

Another important notion of distance that enters into our discussion, which however is not a metric, is the **Kullback-Leibler divergence** (also known as **relative entropy**) between PDFs or positive measures in general. This is given by $D_{\rm KL}(\mathrm{d}\rho_1\|\mathrm{d}\rho_2) := \int(\frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2})\log(\frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2})\mathrm{d}\rho_2$ where $\frac{\mathrm{d}\rho_1}{\mathrm{d}\rho_2}$ denotes the Radon-Nikodym derivative. When[1] $\mathrm{d}\rho_i = \rho_i(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, $i \in \{1,2\}$, are absolutely continuous with respect to the Lebesgue measure, then

$$D_{\rm KL}(\mathrm{d}\rho_1\|\mathrm{d}\rho_2) = \int_{\mathbb{R}^n}\rho_1(\boldsymbol{x})\log\frac{\rho_1(\boldsymbol{x})}{\rho_2(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}.$$

Gradient flow requires an **energy functional**, which we denote by $\mathcal{E}(\rho) := \int U(\boldsymbol{x})\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, where $U(\cdot)$ is the potential energy. Then, a stochastically driven gradient flow is modeled by the Itô SDE (4a) and the Fokker-Planck equation (9) for the corresponding PDF as before. The stationary solution of (9) is the Gibbs distribution $\rho_\infty(\boldsymbol{x}) = \frac{1}{Z}e^{-\beta U(\boldsymbol{x})}$, where the normalization constant $Z := \int_{\mathbb{R}^n}e^{-\beta U(\boldsymbol{x})}\mathrm{d}\boldsymbol{x}$ is known as the **partition function**. The distance to equilibrium which, in a way, quantifies the amount of work that the system can deliver, is captured by the so-called **free energy functional** $\mathcal{F}(\rho)$, defined as the sum of the energy functional $\mathcal{E}(\rho)$ and the negative differential entropy $\mathcal{S}(\rho)$ given in (3), that is,

$$\mathcal{F}(\rho) := \mathcal{E}(\rho) + \beta^{-1}\mathcal{S}(\rho) \tag{11a}$$

$$= \beta^{-1}D_{\rm KL}\left(\rho\|e^{-\beta U(\boldsymbol{x})}\right). \tag{11b}$$

[1]Here we use a slight abuse of notation in that we denote both, the measure and the density with the same symbol.

For the case of (4a), the JKO scheme becomes

$$\varrho_k(\boldsymbol{x},h) = \underset{\varrho\in\mathscr{D}_2}{\operatorname{arginf}}\{\frac{1}{2}W_2^2(\varrho,\varrho_{k-1}) + h\,\mathcal{F}(\varrho)\},\ k\in\mathbb{N}, \tag{12}$$

for step-size $h>0$, and initialized by a given ϱ_0 (satisfying $\mathcal{F}(\varrho_0)<\infty$). For $U(\boldsymbol{x})\equiv 0$, (12) reduces to (2). Solving (12) results in a sequence of PDFs $\{\varrho_k(\boldsymbol{x},h)\}_{k\in\mathbb{N}}$ in $\mathscr{D}_2$. It can be shown following [1] that $\varrho_k(\boldsymbol{x},h)\rightharpoonup\rho(\boldsymbol{x}(t),t)$ weakly in $L^1(\mathbb{R}^n)$ for $t\in[(k-1)h,kh)$, $k\in\mathbb{N}$, as $h\downarrow 0$.

III. JKO Scheme for linear Gaussian systems

We now develop and solve the JKO scheme for the linear Gaussian system in (8) with $\rho_0=\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{P}_0)$ and $\boldsymbol{Q}\equiv\boldsymbol{I}$, without loss of generality. Therefore, we are concerned with the linear Fokker-Planck (Kolmogorov's forward) PDE

$$\frac{\partial\rho}{\partial t} = -\nabla\cdot(\rho\boldsymbol{A}\boldsymbol{x}) + \frac{1}{2}\sum_{i,j=1}^{n}\frac{\partial^2}{\partial x_i\partial x_j}\left(\rho\boldsymbol{B}\boldsymbol{Q}\boldsymbol{B}^\top\right)_{ij}. \tag{13}$$

Under the stated assumptions, it is well-known that (13) admits a steady-state, which is Gaussian with mean zero and covariance $\boldsymbol{P}_\infty\succ\boldsymbol{0}$ that uniquely solves the algebraic Lyapunov equation $\boldsymbol{A}\boldsymbol{P}_\infty+\boldsymbol{P}_\infty\boldsymbol{A}^\top+\boldsymbol{B}\boldsymbol{Q}\boldsymbol{B}^\top=\boldsymbol{0}$. Also, starting from $\rho_0(\boldsymbol{x})=\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{P}_0)$, the transient is $\rho(\boldsymbol{x}(t),t)=\mathcal{N}(\boldsymbol{\mu}(t),\boldsymbol{P}(t))$ where the $\boldsymbol{\mu}(t)$ and $\boldsymbol{P}(t)$ satisfy the following ordinary differential equations (ODEs)

$$\dot{\boldsymbol{\mu}}(t)=\boldsymbol{A}\boldsymbol{\mu}(t),\ \boldsymbol{\mu}(0)=\boldsymbol{\mu}_0, \tag{14a}$$

$$\dot{\boldsymbol{P}}(t)=\boldsymbol{A}\boldsymbol{P}(t)+\boldsymbol{P}(t)\boldsymbol{A}^\top+\boldsymbol{B}\boldsymbol{Q}\boldsymbol{B}^\top,\ \boldsymbol{P}(0)=\boldsymbol{P}_0. \tag{14b}$$

Below, we recover these equations using the JKO scheme. First, in Section III-A, we explain how this is done when $\boldsymbol{A}$ is symmetric and $\boldsymbol{B}\equiv\sqrt{2\beta^{-1}}\boldsymbol{I}$, $\beta>0$, in which case, $\boldsymbol{A}\boldsymbol{x}=-\nabla U(\boldsymbol{x})$ for a suitable potential. The general case, in Section III-B, is more involved and requires to view the drift as the gradient of a time-varying potential.

A. The case where $\boldsymbol{A}$ is symmetric and $\boldsymbol{B}\equiv\sqrt{2\beta^{-1}}\boldsymbol{I}$

Since $\boldsymbol{B}\equiv\sqrt{2\beta^{-1}}\boldsymbol{I}$, $(\boldsymbol{A},\boldsymbol{B})$ is a controllable pair. Further, since $\boldsymbol{A}$ is Hurwitz and symmetric, $\boldsymbol{\Gamma}:=-\boldsymbol{A}\succ 0$, and utilizing the potential

$$U(\boldsymbol{x}):=\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Gamma}\boldsymbol{x}\geq 0,$$

we can cast (8) in the canonical form (4a). Then,

$$\mathcal{E}(\varrho):=\mathbb{E}[U(\boldsymbol{x})]=\frac{1}{2}\left(\boldsymbol{\mu}^\top\boldsymbol{\Gamma}\boldsymbol{\mu}+\operatorname{tr}(\boldsymbol{\Gamma}\boldsymbol{P})\right),$$

where $\boldsymbol{P}$ is the covariance of $\boldsymbol{x}$. Notice that $\mathcal{E}(\cdot)$ depends on the PDF of $\boldsymbol{x}$ *only via its mean and covariance*.

To carry out the optimization (12) over $\mathscr{D}_2$, we adopt a **two-step strategy**. Our approach is motivated by the observation that the objective function in (12) is a sum of two functionals. In the **first step**, we choose a *suitable* parameterized subset of $\mathscr{D}_2$ in such a way that when we optimize the functionals $\frac{1}{2}W^2(\varrho,\varrho_0)$ and $h\mathcal{F}(\varrho)$ *individually over this chosen subspace*, the arginfs (which are achieved) of the two individual optimization problems match. Hence, the sum of the two has the same arginf over the chosen subspace. In the **second step**, we optimize over the subspace parameters. Our choice for the parameterized set of densities is $\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}\subset\mathscr{D}_2$, i.e., the PDFs with given mean-covariance pair $(\boldsymbol{\mu},\boldsymbol{P})$; the choice of the optimal pair is to be decided in the second optimization step.

The development below requires several technical lemmas that are collected in the Appendix.

1) Optimizing over $\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}$: Given $\varrho_0\equiv\rho_0=\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{P}_0)$, and a $\boldsymbol{\mu}$ and $\boldsymbol{P}\succ\boldsymbol{0}$, we first determine

$$\varrho_1=\underset{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}}{\operatorname{arginf}}\{\frac{1}{2}W_2^2(\varrho,\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{P}_0)) + h\,\mathcal{F}(\varrho)\}. \tag{15}$$

From Lemma 2 we see that $\underset{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}}{\operatorname{arginf}}\frac{1}{2}W_2^2(\varrho,\mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{P}_0))$ is achieved by $\varrho=\mathcal{N}(\boldsymbol{\mu},\boldsymbol{P})$ (uniquely). From Lemma 3, since $U(\boldsymbol{x})=\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Gamma}\boldsymbol{x}$, we also know that $\underset{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}}{\operatorname{arginf}}h\,\mathcal{F}(\varrho)$ is achieved by $\varrho=\mathcal{N}(\boldsymbol{\mu},\boldsymbol{P})$ (uniquely). Thus, $\varrho_1=\mathcal{N}(\boldsymbol{\mu},\boldsymbol{P})$. The infimal value in (15) is now the sum of the two infima,

$$\frac{1}{2}\left[\|\boldsymbol{\mu}-\boldsymbol{\mu}_0\|_2^2+\operatorname{tr}\left(\boldsymbol{P}+\boldsymbol{P}_0-2\left(\boldsymbol{P}_0^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{P}_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)\right]+\frac{h}{2\beta}\left[-n-n\log(2\pi)-\log\det(\boldsymbol{P})+\beta\boldsymbol{\mu}^\top\boldsymbol{\Gamma}\boldsymbol{\mu}+\beta\operatorname{tr}(\boldsymbol{\Gamma}\boldsymbol{P})\right]. \tag{16}$$

2) Optimizing over $(\boldsymbol{\mu},\boldsymbol{P})$: Equating the gradient of (16) w.r.t. $\boldsymbol{\mu}$ to zero, results $\boldsymbol{\mu}=\boldsymbol{\phi}(\boldsymbol{\mu_0}):=(\boldsymbol{I}+h\boldsymbol{\Gamma})^{-1}\boldsymbol{\mu}_0$. The recursion $\boldsymbol{\mu}_k=\boldsymbol{\phi}(\boldsymbol{\mu_{k-1}})$, up to first order in h, becomes

$$\boldsymbol{\mu}_k=(\boldsymbol{I}-h\boldsymbol{\Gamma})\,\boldsymbol{\mu}_{k-1}+O(h^2). \tag{17}$$

We see that this recursion coincides with the solution of (14a) in the "small h" limit. Specifically, $\boldsymbol{\mu}(t)=e^{\boldsymbol{A}t}\boldsymbol{\mu}_0\Rightarrow\boldsymbol{\mu}_k:=\boldsymbol{\mu}(t=kh)=\left(e^{\boldsymbol{A}h}\right)^k\boldsymbol{\mu}_0\Rightarrow\boldsymbol{\mu}_k=e^{\boldsymbol{A}h}\boldsymbol{\mu}_{k-1}=(\boldsymbol{I}+h\boldsymbol{A})\,\boldsymbol{\mu}_{k-1}+O(h^2)$, which is same as (17) since $\boldsymbol{\Gamma}:=-\boldsymbol{A}$. Thus, we have recovered (14a) using discrete time-stepping via JKO scheme in the small step-size limit.

Setting the gradient of (16) w.r.t. $\boldsymbol{P}$ to zero (using Lemma 4), we obtain

$$\boldsymbol{I}-\boldsymbol{P}_0^{\frac{1}{2}}\left(\boldsymbol{P}_0^{-\frac{1}{2}}\boldsymbol{P}^{-1}\boldsymbol{P}_0^{-\frac{1}{2}}\right)^{\frac{1}{2}}\boldsymbol{P}_0^{\frac{1}{2}}-h\beta\boldsymbol{P}^{-1}+h\boldsymbol{\Gamma}=\boldsymbol{0}. \tag{18}$$

By pre and post multiplying both sides of (18) with $\boldsymbol{P}_0^{-\frac{1}{2}}$, and letting $\left(\boldsymbol{P}_0^{-\frac{1}{2}}\boldsymbol{P}^{-1}\boldsymbol{P}_0^{-\frac{1}{2}}\right)^{\frac{1}{2}}=:\boldsymbol{Z}$, we arrive at

$$\boldsymbol{Z}^2+\frac{\beta}{h}\boldsymbol{Z}-\frac{\beta}{h}\boldsymbol{P}_0^{-\frac{1}{2}}(\boldsymbol{I}+h\boldsymbol{\Gamma})\boldsymbol{P}_0^{-\frac{1}{2}}=\boldsymbol{0},$$

which admits the unique closed-form solution [13, p. 304]

$$\boldsymbol{Z}=\frac{\beta}{2h}\left(-\boldsymbol{I}+\left(\boldsymbol{I}+4\frac{h}{\beta}\boldsymbol{P}_0^{-\frac{1}{2}}(\boldsymbol{I}+h\boldsymbol{\Gamma})\,\boldsymbol{P}_0^{-\frac{1}{2}}\right)^{\frac{1}{2}}\right). \tag{19}$$

Expanding (19), we obtain

$$\boldsymbol{Z} = \frac{\beta}{2h}\Bigg[-\boldsymbol{I} + \Bigg\{\boldsymbol{I} + \frac{1}{2}4\frac{h}{\beta}\boldsymbol{P}_0^{-\frac{1}{2}}\left(\boldsymbol{I}+h\boldsymbol{\Gamma}\right)\boldsymbol{P}_0^{-\frac{1}{2}} + \frac{\frac{1}{2}\left(\frac{1}{2}-1\right)}{2!}\frac{16h^2}{\beta^2}\boldsymbol{P}_0^{-\frac{1}{2}}\left(\boldsymbol{I}+h\boldsymbol{\Gamma}\right)\boldsymbol{P}_0^{-1}\left(\boldsymbol{I}+h\boldsymbol{\Gamma}\right)\boldsymbol{P}_0^{-\frac{1}{2}} + O(h^3)\Bigg\}\Bigg]$$
$$= \boldsymbol{P}_0^{-\frac{1}{2}}\left(\boldsymbol{I} + h\boldsymbol{\Gamma} - \frac{h}{\beta}\boldsymbol{P}_0^{-1}\right)\boldsymbol{P}_0^{-\frac{1}{2}} \;+\; O(h^2). \tag{20}$$

Substituting $\boldsymbol{Z} = \left(\boldsymbol{P}_0^{-\frac{1}{2}}\boldsymbol{P}^{-1}\boldsymbol{P}_0^{-\frac{1}{2}}\right)^{\frac{1}{2}}$ back into (20), squaring, and rearranging, we get that

$$\boldsymbol{P} = \left(\boldsymbol{I} + h\left(\boldsymbol{\Gamma} - \frac{1}{\beta}\boldsymbol{P}_0^{-1}\right)\right)^{-1}\boldsymbol{P}_0\left(\boldsymbol{I} + h\left(\boldsymbol{\Gamma} - \frac{1}{\beta}\boldsymbol{P}_0^{-1}\right)\right)^{-1} + O(h^2)$$
$$= \left(\boldsymbol{I} - h\left(\boldsymbol{\Gamma} - \frac{1}{\beta}\boldsymbol{P}_0^{-1}\right)\right)\boldsymbol{P}_0\left(\boldsymbol{I} - h\left(\boldsymbol{\Gamma} - \frac{1}{\beta}\boldsymbol{P}_0^{-1}\right)\right) \;+\; O(h^2)$$
$$= \boldsymbol{\Psi}(\boldsymbol{P}_0) \;+\; O(h^2),$$

where $\boldsymbol{\Psi}(\boldsymbol{P}_0) := \boldsymbol{P}_0 + h\left(-\boldsymbol{\Gamma}\boldsymbol{P}_0 - \boldsymbol{P}_0\boldsymbol{\Gamma} + 2\beta^{-1}\boldsymbol{I}\right)$. Set the matrix-valued recursion $\boldsymbol{P}_k = \boldsymbol{\Psi}\left(\boldsymbol{P}_{k-1}\right)$, where

$$\boldsymbol{\Psi}\left(\boldsymbol{P}_{k-1}\right) := \boldsymbol{P}_{k-1} + h\left(-\boldsymbol{\Gamma}\boldsymbol{P}_0 - \boldsymbol{P}_0\boldsymbol{\Gamma} + 2\beta^{-1}\boldsymbol{I}\right) + O(h^2). \tag{21}$$

To show that (21) indeed recovers (14b), first notice that substituting $\boldsymbol{A} = \boldsymbol{A}^\top = -\boldsymbol{\Gamma}$ and $\boldsymbol{B} = \sqrt{2\beta^{-1}}\boldsymbol{I}$ in (14b) results the Lyapunov differential equation

$$\dot{\boldsymbol{P}}(t) = -\boldsymbol{\Gamma}\boldsymbol{P}(t) - \boldsymbol{P}(t)\boldsymbol{\Gamma} + 2\beta^{-1}\boldsymbol{I}$$

subject to $\boldsymbol{P}(0) = \boldsymbol{P}_0$, which can be solved via the method of integrating factor as

$$\boldsymbol{P}(t) = \frac{1}{\beta}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{I} - e^{-2\boldsymbol{\Gamma}t}\right) \;+\; e^{-\boldsymbol{\Gamma}t}\boldsymbol{P}_0 e^{-\boldsymbol{\Gamma}t}. \tag{22}$$

Thus, for $t = kh$,

$$\boldsymbol{P}_k := \boldsymbol{P}(kh) = \beta^{-1}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{I} - e^{-2\boldsymbol{\Gamma}kh}\right) + e^{-\boldsymbol{\Gamma}kh}\boldsymbol{P}_0 e^{-\boldsymbol{\Gamma}kh}$$
$$= 2\beta^{-1}kh\boldsymbol{I} + (\boldsymbol{P}_0 - kh\boldsymbol{\Gamma}\boldsymbol{P}_0 - kh\boldsymbol{P}_0\boldsymbol{\Gamma}) + O(h^2).$$

Replacing k with $k-1$ in the latter yields a similar expression for $\boldsymbol{P}_{k-1}$. Then, subtracting these expressions for $\boldsymbol{P}_{k-1}$ from $\boldsymbol{P}_k$ we obtain that

$$\boldsymbol{P}_k - \boldsymbol{P}_{k-1} = 2\beta^{-1}h\boldsymbol{I} - h\boldsymbol{\Gamma}\boldsymbol{P}_0 - h\boldsymbol{P}_0\boldsymbol{\Gamma} + O(h^2),$$

which is same as (21) derived from JKO scheme. Thus, we have recovered the covariance evolution through Fokker-Planck dynamics using the time-stepping procedure via JKO scheme in the small step-size limit.

B. The case of Hurwitz $\boldsymbol{A}$ and controllable $(\boldsymbol{A}, \boldsymbol{B})$

We scale $\boldsymbol{B}$ into $\sqrt{2}\boldsymbol{B}$ without loss of generality, and take as initial PDF $\varrho_0 \equiv \rho_0 = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{P}_0)$. Since we allow any Hurwitz (not necessarily symmetric) $\boldsymbol{A}$, and any $\boldsymbol{B}$ that makes $(\boldsymbol{A}, \sqrt{2}\boldsymbol{B})$ a controllable pair, it is not apparent if and how one can express (8) in the canonical form (4a). The main impediment in doing so, is twofold: (1) how to define the potential energy $U(\boldsymbol{x})$, and (2) how to interpret and define the parameter β in the generic case. In the following, we show that by two successive co-ordinate transformations, system (8) can indeed be put in the form (4a). Similar transformations have been mentioned in [14, p. 1464], [15] in a different context.

1) Equipartition of energy coordinate transformation: Consider the stationary covariance $\boldsymbol{P}_\infty$ associated with $(\boldsymbol{A}, \sqrt{2}\boldsymbol{B})$ that satisfies

$$\boldsymbol{A}\boldsymbol{P}_\infty + \boldsymbol{P}_\infty\boldsymbol{A}^\top + 2\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{0}. \tag{23}$$

For a system at a stationary distribution, we define the **thermodynamic temperature** θ as the average amount of "energy" per degree of freedom, that is,

$$\theta := \frac{1}{n}\operatorname{tr}(\boldsymbol{P}_\infty),$$

and, thereby, $\beta := \theta^{-1}$ the **inverse temperature**. By pre and post multiplying (23) with $\boldsymbol{P}_\infty^{-\frac{1}{2}}$, and rescaling by θ so as to preserve the temperature in the new coordinates, we get

$$\boldsymbol{A}_{\rm ep}\theta\boldsymbol{I} + \theta\boldsymbol{I}\boldsymbol{A}_{\rm ep}^\top + \sqrt{2\theta}\boldsymbol{B}_{\rm ep}(\sqrt{2\theta}\boldsymbol{B}_{\rm ep})^\top = \boldsymbol{0}, \tag{24}$$

where $\boldsymbol{A}_{\rm ep} := \boldsymbol{P}_\infty^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{P}_\infty^{\frac{1}{2}}$, $\boldsymbol{B}_{\rm ep} := \boldsymbol{P}_\infty^{-\frac{1}{2}}\boldsymbol{B}$, while the stationary covariance $\theta\boldsymbol{I}$ reflects *equipartition of energy*. The equipartition of energy co-ordinate transformation $(\boldsymbol{A}, \sqrt{2}\boldsymbol{B}) \mapsto (\boldsymbol{A}_{\rm ep}, \sqrt{2\theta}\boldsymbol{B}_{\rm ep})$, corresponds to the state-transformation $\boldsymbol{x} \mapsto \boldsymbol{x}_{\rm ep} := \sqrt{\theta}\boldsymbol{P}_\infty^{-\frac{1}{2}}\boldsymbol{x}$, leading to

$$\mathrm{d}\boldsymbol{x}_{\rm ep}(t) = \boldsymbol{A}_{\rm ep}\boldsymbol{x}_{\rm ep}(t)\,\mathrm{d}t \;+\; \sqrt{2\theta}\boldsymbol{B}_{\rm ep}\,\mathrm{d}\boldsymbol{w}(t). \tag{25}$$

This settles how β is to be defined and interpreted in the context of JKO scheme (11) and (12). On the other hand, $\boldsymbol{A}_{\rm ep}$ being similar to $\boldsymbol{A}$, is guaranteed to be Hurwitz but not symmetric, unless $\boldsymbol{A}$ was symmetric to begin with. Thus, it remains for us to "symmetrize" $\boldsymbol{A}_{\rm ep}$ and define a suitable potential energy $U(\cdot)$ as needed in (12). We do this next.

2) Symmetrization transformation: We introduce the time-varying transformation

$$\boldsymbol{x}_{\rm ep} \mapsto \boldsymbol{x}_{\rm sym} := e^{-\boldsymbol{A}_{\rm ep}^{\rm skew}t}\boldsymbol{x}_{\rm ep}$$

where $\boldsymbol{A}_{\rm ep}^{\rm skew} := \frac{1}{2}(\boldsymbol{A}_{\rm ep} - \boldsymbol{A}_{\rm ep}^\top)$. This results in

$$(\boldsymbol{A}_{\rm ep}, \sqrt{2\theta}\boldsymbol{B}_{\rm ep}) \mapsto (\boldsymbol{F}(t), \sqrt{2\theta}\boldsymbol{G}(t)),$$

with

$$\boldsymbol{F}(t) := e^{-\boldsymbol{A}_{\rm ep}^{\rm skew}t}\boldsymbol{A}_{\rm ep}^{\rm sym}e^{\boldsymbol{A}_{\rm ep}^{\rm skew}t}, \text{ and } \boldsymbol{G}(t) := e^{-\boldsymbol{A}_{\rm ep}^{\rm skew}t}\boldsymbol{B}_{\rm ep},$$

where, similarly, $\boldsymbol{A}_{\rm ep}^{\rm sym} := \frac{1}{2}(\boldsymbol{A}_{\rm ep} + \boldsymbol{A}_{\rm ep}^\top)$. Thus, $\boldsymbol{x}_{\rm sym}(t)$ satisfies

$$\mathrm{d}\boldsymbol{x}_{\rm sym}(t) = \boldsymbol{F}(t)\boldsymbol{x}_{\rm sym}(t)\,\mathrm{d}t \;+\; \sqrt{2\theta}\boldsymbol{G}(t)\,\mathrm{d}\boldsymbol{w}(t). \tag{26}$$

Notice that $\boldsymbol{F}(t)$ is symmetric for all t. Furthermore, observe that the new coordinates $\boldsymbol{x}_{\rm sym}$ is simply obtained by a (time-varying) orthogonal transformation of the equipartition of energy coordinates $\boldsymbol{x}_{\rm ep}$. Hence the stationary covariance of $\boldsymbol{x}_{\rm sym}$ is identical to that of $\boldsymbol{x}_{\rm ep}$, which is $\theta\boldsymbol{I}$ (from Section III-B.1). What happens is that the covariance $\boldsymbol{x}_{\rm sym}(t)$ tends to the same steady state value as $t \to \infty$ *in spite of the fact that* (26) *has time varying coefficients*. To see this in different way, we can rewrite (24) as $\boldsymbol{B}_{\rm ep}\boldsymbol{B}_{\rm ep}^\top = -\boldsymbol{A}_{\rm ep}^{\rm sym}$, and deduce that

$$\boldsymbol{G}(t)\boldsymbol{G}(t)^\top = e^{-\boldsymbol{A}_{\rm ep}^{\rm skew}t}\boldsymbol{B}_{\rm ep}\boldsymbol{B}_{\rm ep}^\top e^{\boldsymbol{A}_{\rm ep}^{\rm skew}t} = -\boldsymbol{F}(t),$$
$$\Rightarrow \boldsymbol{F}(t)\theta\boldsymbol{I} + \theta\boldsymbol{I}\boldsymbol{F}(t) + \sqrt{2\theta}\boldsymbol{G}(t)(\sqrt{2\theta}\boldsymbol{G}(t))^\top = \boldsymbol{0}. \tag{27}$$

The symmetrization $\boldsymbol{x}_{\rm ep} \mapsto \boldsymbol{x}_{\rm sym}$ leaves the stationary covariance $\theta\boldsymbol{I}$ invariant. This guarantees the definition of temperature θ stays intact.

3) Recovery of the Fokker-Planck solution: We are now ready to apply the JKO scheme to the generic stochastic linear system $\mathrm{d}\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{x}(t)\,\mathrm{d}t + \sqrt{2}\boldsymbol{B}\,\mathrm{d}\boldsymbol{w}(t)$, with initial PDF $\rho(\boldsymbol{x}(0),0) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{P}_0)$. To this end, we carry out a computation akin to the two steps in Section III-A, for the transformed SDE (26) in the symmetrized coordinate $\boldsymbol{x}_{\rm sym}$. From there on, we recover the Fokker-Planck solution in the original coordinate $\boldsymbol{x}$.

Since $\boldsymbol{x} \mapsto \boldsymbol{x}_{\rm sym}$ is a linear transformation, it follows that $\boldsymbol{x}_{\rm sym} \sim \mathcal{N}(\boldsymbol{\mu}_{\rm sym}, \boldsymbol{P}_{\rm sym})$ whenever $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{P})$. Thus, carrying out the first step of the optimization in $\boldsymbol{x}_{\rm sym}$ coordinate, we get an expression similar to (16) wherein $(\boldsymbol{\mu}, \boldsymbol{P})$ is to be replaced by $(\boldsymbol{\mu}_{\rm sym}, \boldsymbol{P}_{\rm sym})$, and $(\boldsymbol{\mu}_0, \boldsymbol{P}_0)$ is to be replaced by $(\boldsymbol{\mu}_{{\rm sym}_0}, \boldsymbol{P}_{{\rm sym}_0})$. To carry out the second step of optimization, notice that $\boldsymbol{A}_{\rm ep}^{\rm sym} = -\boldsymbol{B}_{\rm ep}\boldsymbol{B}_{\rm ep}^\top \preceq 0$, and consequently $\boldsymbol{F}(t) = e^{-\boldsymbol{A}_{\rm ep}^{\rm skew}t}\boldsymbol{A}_{\rm ep}^{\rm sym}e^{\boldsymbol{A}_{\rm ep}^{\rm skew}t} \preceq 0$. Thus, considering the time-varying potential

$$U(\boldsymbol{x}_{\rm sym}) := -\frac{1}{2}\boldsymbol{x}_{\rm sym}^\top \boldsymbol{F}(t)\boldsymbol{x}_{\rm sym} \geq 0,$$

and setting the partial derivative of the infimal value from first stage of the optimization w.r.t. $\boldsymbol{\mu}_{\rm sym}$ to zero, results the recursion $\boldsymbol{\mu}_{{\rm sym}_k} = (\boldsymbol{I} - h\boldsymbol{F}(kh))^{-1}\boldsymbol{\mu}_{{\rm sym}_{k-1}}$. Recalling that $\boldsymbol{x}_{\rm sym} = e^{-\boldsymbol{A}_{\rm ep}^{\rm skew}t}\sqrt{\theta}\boldsymbol{P}_\infty^{-\frac{1}{2}}\boldsymbol{x}$, we arrive at a recursion in original coordinate:

$$\boldsymbol{\mu}_k = \boldsymbol{P}_\infty^{\frac{1}{2}} e^{\boldsymbol{A}_{\rm ep}^{\rm skew}kh}\{(\boldsymbol{I} - h\boldsymbol{F}(kh))^{-1}e^{\boldsymbol{A}_{\rm ep}^{\rm skew}h}\} e^{-\boldsymbol{A}_{\rm ep}^{\rm skew}kh}\boldsymbol{P}_\infty^{-\frac{1}{2}}\,\boldsymbol{\mu}_{k-1}. \tag{28}$$

By a series expansion and collecting linear terms in h, one can verify the following:

$$\begin{aligned}
&(\boldsymbol{I} - h\boldsymbol{F}(kh))^{-1} = \boldsymbol{I} + h\boldsymbol{A}_{\rm ep}^{\rm sym} + O(h^2),\\
&(\boldsymbol{I} - h\boldsymbol{F}(kh))^{-1}e^{\boldsymbol{A}_{\rm ep}^{\rm skew}h} = \boldsymbol{I} + h\boldsymbol{A}_{\rm ep} + O(h^2),\\
&e^{\boldsymbol{A}_{\rm ep}^{\rm skew}kh}(\boldsymbol{I} - h\boldsymbol{F}(kh))^{-1}e^{\boldsymbol{A}_{\rm ep}^{\rm skew}h}e^{-\boldsymbol{A}_{\rm ep}^{\rm skew}kh}\\
&\qquad = \boldsymbol{I} + h\boldsymbol{A}_{\rm ep} + O(h^2).
\end{aligned}$$

Hence (28) yields

$$\begin{aligned}
\boldsymbol{\mu}_k &= \left(\boldsymbol{I} + h\boldsymbol{P}_\infty^{\frac{1}{2}}\boldsymbol{A}_{\rm ep}\boldsymbol{P}_\infty^{-\frac{1}{2}}\right)\boldsymbol{\mu}_{k-1} \;+\; O(h^2)\\
&= (\boldsymbol{I} + h\boldsymbol{A})\,\boldsymbol{\mu}_{k-1} \;+\; O(h^2),
\end{aligned} \tag{29}$$

where the last equality follows from $\boldsymbol{A}_{\rm ep} := \boldsymbol{P}_\infty^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{P}_\infty^{\frac{1}{2}}$. Since $\dot{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{\mu}$ and $e^{h\boldsymbol{A}} = \boldsymbol{I} + h\boldsymbol{A} + O(h^2)$, in the small h limit, equation (29) thus recovers (14a), as in Section III-A. A similar straightforward but tedious computation leads to the matrix recursion

$$\boldsymbol{P}_k - \boldsymbol{P}_{k-1} = h(\boldsymbol{A}\boldsymbol{P}_{k-1} + \boldsymbol{P}_{k-1}\boldsymbol{A}^\top + 2\boldsymbol{B}\boldsymbol{B}^\top) + O(h^2), \tag{30}$$

which in the limit $h \downarrow 0$, is indeed a first-order approximation of the Lyapunov equation for the original system. We omit the details for brevity.

IV. JKO-like Schemes for Filtering

In this section, we focus on the linear Gaussian filtering problem, with process model and measurement models

$$\begin{aligned}
\mathrm{d}\boldsymbol{x}(t) &= \boldsymbol{A}\boldsymbol{x}(t)\mathrm{d}t + \sqrt{2}\boldsymbol{B}\mathrm{d}\boldsymbol{w}(t)\\
\mathrm{d}\boldsymbol{z}(t) &= \boldsymbol{C}\boldsymbol{x}(t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{v}(t),
\end{aligned}$$

where $\boldsymbol{C} \in \mathbb{R}^{m\times n}$, and $\rho_0 = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{P}_0)$. The conditional PDF $\rho^+(\boldsymbol{x}(t),t) = \mathcal{N}(\boldsymbol{\mu}^+(t), \boldsymbol{P}^+(t))$, given measurements up to time t, is well-known and given by the **Kalman-Bucy filter** [20]

$$\mathrm{d}\boldsymbol{\mu}^+(t) = \boldsymbol{A}\boldsymbol{\mu}^+(t)\mathrm{d}t + \boldsymbol{K}(t)\left(\mathrm{d}\boldsymbol{z}(t) - \boldsymbol{C}\boldsymbol{\mu}^+(t)\mathrm{d}t\right), \tag{31a}$$

$$\dot{\boldsymbol{P}}^+(t) = \boldsymbol{A}\boldsymbol{P}^+(t) + \boldsymbol{P}^+(t)\boldsymbol{A}^\top + 2\boldsymbol{B}\boldsymbol{B}^\top - \boldsymbol{K}(t)\boldsymbol{R}\boldsymbol{K}(t)^\top \tag{31b}$$

that specifies a vector SDE and a matrix ODE, respectively, for the conditional mean $\boldsymbol{\mu}^+(t)$ and covariance $\boldsymbol{P}^+(t)$. The initial conditions are $\boldsymbol{\mu}^+(0) = \boldsymbol{\mu}_0$, $\boldsymbol{P}^+(0) = \boldsymbol{P}_0$, and $\boldsymbol{K}(t) := \boldsymbol{P}^+(t)\boldsymbol{C}^\top\boldsymbol{R}^{-1}$ is the so-called Kalman gain.

In the sequel, we demonstrate that by applying the two-step optimization strategy we used before in Section III, we can recover the Kalman-Bucy filter from LMMR-equation (5) for the linear Gaussian case as the $h \downarrow 0$ limit.

A. LMMR gradient descent scheme

Once again we proceed with carrying out the following two optimization steps. First, we optimize (5) over $\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}$, and then optimize the minimum value over the choice of parameters $(\boldsymbol{\mu}, \boldsymbol{P})$.

1) Optimizing over $\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}$: Consider $\varrho_k^- = \mathcal{N}(\boldsymbol{\mu}_k^-, \boldsymbol{P}_k^-)$ to be our prior for the state PDF at time $t = kh$. Observe that

$$\inf_{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}} D_{\rm KL}\left(\varrho \| \mathcal{N}(\boldsymbol{\mu}_k^-, \boldsymbol{P}_k^-)\right) = \inf_{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}}\left[\int_{\mathbb{R}^n}\varrho(\boldsymbol{x})\log\varrho(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \mathbb{E}_\varrho\{\log\mathcal{N}(\boldsymbol{\mu}_k^-, \boldsymbol{P}_k^-)\}\right]. \tag{32}$$

and that

$$\mathbb{E}_\varrho\{\log\mathcal{N}(\boldsymbol{\mu}_k^-, \boldsymbol{P}_k^-)\} = -\frac{1}{2}\left[(\boldsymbol{\mu} - \boldsymbol{\mu}_k^-)^\top\left(\boldsymbol{P}_k^-\right)^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_k^-) + \mathrm{tr}\left(\boldsymbol{P}(\boldsymbol{P}_k^-)^{-1}\right)\right] - \frac{1}{2}\log\left((2\pi)^n\det(\boldsymbol{P}_k^-)\right)$$

remains invariant for all $\varrho \in \mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}$. Therefore, the arginf in (32) is achieved by the Gaussian PDF $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{P})$ (i.e., the maximum entropy PDF with given mean-covariance), and the infimal value is precisely $D_{\rm KL}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{P})\|\mathcal{N}(\boldsymbol{\mu}_k^-, \boldsymbol{P}_k^-))$. On the other hand, notice that

$$\inf_{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}}\frac{1}{2}\,\mathbb{E}_\varrho\{(\boldsymbol{y}_k - \boldsymbol{C}\boldsymbol{x})^\top\boldsymbol{R}^{-1}(\boldsymbol{y}_k - \boldsymbol{C}\boldsymbol{x})\} = \frac{1}{2}\left[(\boldsymbol{y}_k - \boldsymbol{C}\boldsymbol{\mu})^\top \boldsymbol{R}^{-1}(\boldsymbol{y}_k - \boldsymbol{C}\boldsymbol{\mu}) + \mathrm{tr}\left(\boldsymbol{C}^\top\boldsymbol{R}^{-1}\boldsymbol{C}\boldsymbol{P}\right)\right] = \text{constant} \tag{33}$$

Attribute ↓ \ Coordinate →	Original	Equipartition of energy	Symmetrization
State vector	$\boldsymbol{x}$	$\boldsymbol{x}_{\rm ep}$	$\boldsymbol{x}_{\rm sym}$
System matrices	$(\boldsymbol{A}, \sqrt{2}\boldsymbol{B})$	$(\boldsymbol{A}_{\rm ep}, \sqrt{2\theta}\boldsymbol{B}_{\rm ep})$	$(\boldsymbol{F}(t), \sqrt{2\theta}\boldsymbol{G}(t))$
Stationary covariance	$\boldsymbol{P}_\infty$	$\theta\boldsymbol{I}$	$\theta\boldsymbol{I}$

TABLE I: Summary of the coordinate transformations for Section III-B.

as well over $\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}$. Hence

$$\underset{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}}{\operatorname{arginf}} \Big[D_{\rm KL}\left(\varrho\|\mathcal{N}(\boldsymbol{\mu}_k^-,\boldsymbol{P}_k^-)\right) + \frac{h}{2}\,\mathbb{E}_\varrho\{(\boldsymbol{y}_k-\boldsymbol{C}\boldsymbol{x})^\top\boldsymbol{R}^{-1}(\boldsymbol{y}_k-\boldsymbol{C}\boldsymbol{x})\}\Big] = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{P})$$

and the corresponding infimum value is

$$\frac{1}{2}\left[\operatorname{tr}\left((\boldsymbol{P}_k^-)^{-1}\boldsymbol{P}\right) + (\boldsymbol{\mu}_k^- - \boldsymbol{\mu})^\top(\boldsymbol{P}_k^-)^{-1}(\boldsymbol{\mu}_k^- - \boldsymbol{\mu}) - n - \log\det\left((\boldsymbol{P}_k^-)^{-1}\boldsymbol{P}\right)\right] + \frac{h}{2}\left[(\boldsymbol{y}_k-\boldsymbol{C}\boldsymbol{\mu})^\top\boldsymbol{R}^{-1}(\boldsymbol{y}_k-\boldsymbol{C}\boldsymbol{\mu}) + \operatorname{tr}\left(\boldsymbol{C}^\top\boldsymbol{R}^{-1}\boldsymbol{C}\boldsymbol{P}\right)\right]. \quad (34)$$

2) Optimizing over $(\boldsymbol{\mu},\boldsymbol{P})$*:* Equating the partial derivative of (34) w.r.t. $\boldsymbol{\mu}$ to zero, and setting $\boldsymbol{\mu}\equiv\boldsymbol{\mu}_k^+$ in the resulting algebraic equation, we get

$$(\boldsymbol{P}_k^-)^{-1}\left(\boldsymbol{\mu}_k^- - \boldsymbol{\mu}_k^+\right) + h\boldsymbol{C}^\top\boldsymbol{R}^{-1}\left(\boldsymbol{y}_k - \boldsymbol{C}\boldsymbol{\mu}_k^+\right) = \boldsymbol{0},$$
$$\Rightarrow \boldsymbol{\mu}_k^+ = \boldsymbol{\mu}_k^- + h\boldsymbol{P}_k^-\boldsymbol{C}^\top\boldsymbol{R}^{-1}\left(\boldsymbol{y}_k - \boldsymbol{C}\boldsymbol{\mu}_k^+\right). \quad (35)$$

On the other hand, equating the partial derivative of (34) w.r.t. $\boldsymbol{P}$ to zero, and then setting $\boldsymbol{P}\equiv\boldsymbol{P}_k^+$ in the resulting algebraic equation, we get

$$(\boldsymbol{P}_k^+)^{-1} = (\boldsymbol{P}_k^-)^{-1} + h\boldsymbol{C}^\top\boldsymbol{R}^{-1}\boldsymbol{C} \Rightarrow \boldsymbol{P}_k^+ = \left(\boldsymbol{I} + h\boldsymbol{P}_k^-\boldsymbol{C}^\top\boldsymbol{R}^{-1}\boldsymbol{C}\right)^{-1}\boldsymbol{P}_k^- = \boldsymbol{P}_k^- - h\boldsymbol{P}_k^-\boldsymbol{C}^\top\boldsymbol{R}^{-1}\boldsymbol{C}\boldsymbol{P}_k^- + O(h^2). \quad (36)$$

With $\Delta\boldsymbol{z}_k = \boldsymbol{y}_k h$, as in Section I,

$$\mathrm{d}\boldsymbol{z}(t) = \Delta\boldsymbol{z}_k + O(h^2),$$
$$\boldsymbol{\mu}^+(t)\mathrm{d}t = \boldsymbol{\mu}_k^+ h + O(h^2),$$

and from (29) that

$$\boldsymbol{\mu}_k^- = (\boldsymbol{I} + h\boldsymbol{A})\boldsymbol{\mu}_{k-1}^+ + O(h^2).$$

These, together with (36), allow us to simplify (35) as

$$\boldsymbol{\mu}_k^+ - \boldsymbol{\mu}_{k-1}^+ = h\boldsymbol{A}\boldsymbol{\mu}_{k-1}^+ + \boldsymbol{P}_k^+\boldsymbol{C}^\top\boldsymbol{R}^{-1}\left(\Delta\boldsymbol{z}_k - h\boldsymbol{C}\boldsymbol{\mu}_k^+\right) + O(h^2),$$

which in the limit $h\downarrow 0$, leads to (31a).

Substituting (30) into (36) we arrive at

$$\boldsymbol{P}_k^+ - \boldsymbol{P}_{k-1}^+ = h(\boldsymbol{A}\boldsymbol{P}_{k-1}^+ + \boldsymbol{P}_{k-1}^+\boldsymbol{A}^\top + 2\boldsymbol{B}\boldsymbol{B}^\top) - h\boldsymbol{P}_{k-1}^+\boldsymbol{C}^\top\boldsymbol{R}^{-1}\boldsymbol{C}\boldsymbol{P}_{k-1}^+ + O(h^2). \quad (37)$$

In the limit $h\downarrow 0$, (37) recovers (31b).

B. Alternative JKO-like schemes for filtering

The ideas in the LMMR-scheme suggest the possibility of alternative variational schemes to approximate stochastic estimators. Such a viewpoint has been put forth in [21], promoting the notion of regularized dynamic inversion. As an example, one may consider a *gradient descent with respect to the Wasserstein* distance $\frac{1}{2}W_2^2$, instead of KL-divergence $D_{\rm KL}$ in (5). In that case, the posterior may be constructed according to

$$\varrho_k^+(\boldsymbol{x},h) = \underset{\varrho\in\mathcal{D}_2}{\operatorname{arginf}}\ \frac{1}{2}W_2^2\left(\varrho,\varrho_k^-\right) + h\Phi(\varrho), \quad k\in\mathbb{N}, \quad (38)$$

where Φ is as in (6). The template of the two-step optimization again applies and, specializing to the linear Gaussian case, the solution of (38) in the $h\downarrow 0$ limit, is $\mathcal{N}(\boldsymbol{\mu}^+(t),\boldsymbol{P}^+(t))$, given by

$$\mathrm{d}\boldsymbol{\mu}^+(t) = \boldsymbol{A}\boldsymbol{\mu}^+(t)\mathrm{d}t + \boldsymbol{L}\left(\mathrm{d}\boldsymbol{z}(t) - \boldsymbol{C}\boldsymbol{\mu}^+(t)\mathrm{d}t\right),$$
$$\dot{\boldsymbol{P}}^+(t) = (\boldsymbol{A}-\boldsymbol{L}\boldsymbol{C})\boldsymbol{P}^+(t) + \boldsymbol{P}^+(t)(\boldsymbol{A}-\boldsymbol{L}\boldsymbol{C})^\top + 2\boldsymbol{B}\boldsymbol{B}^\top,$$

where $\boldsymbol{L} := \boldsymbol{C}^\top\boldsymbol{R}^{-1}$, and $\boldsymbol{\mu}^+(0)=\boldsymbol{\mu}_0$, $\boldsymbol{P}^+(0)=\boldsymbol{P}_0$. This follows by noticing from Sections III-A.1 and IV-A.1 that

$$\underset{\varrho\in\mathcal{D}_{\boldsymbol{\mu},\boldsymbol{P}}}{\operatorname{arginf}}\left[\frac{1}{2}W_2^2\left(\varrho,\mathcal{N}(\boldsymbol{\mu}_k^-,\boldsymbol{P}_k^-\right) + h\Phi(\varrho)\right] = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{P}),$$

where the infimum value is

$$\frac{1}{2}\left[\|\,\boldsymbol{\mu}-\boldsymbol{\mu}_k^-\,\|_2^2 + \operatorname{tr}\left(\boldsymbol{P}+\boldsymbol{P}_k^- - 2\left((\boldsymbol{P}_k^-)^{\frac{1}{2}}\boldsymbol{P}(\boldsymbol{P}_k^-)^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)\right] + \frac{h}{2}\left[(\boldsymbol{y}_k-\boldsymbol{C}\boldsymbol{\mu})^\top\boldsymbol{R}^{-1}(\boldsymbol{y}_k-\boldsymbol{C}\boldsymbol{\mu}) + \operatorname{tr}\left(\boldsymbol{C}^\top\boldsymbol{R}^{-1}\boldsymbol{C}\boldsymbol{P}\right)\right].$$

It is instructive to compare the above with (31). In this case, the estimator is of a Luenberger type with a static gain matrix $\boldsymbol{L}$ which is decoupled from the covariance, unlike (31). It is obviously not optimal in the minimum mean-square error sense. It is only presented here as a guideline to explore other variational schemes with desirable properties.

V. Concluding remarks

Reformulating uncertainty propagation and the filtering equations as gradient flows [5] is potentially transformative [1] [6]. The full power of this viewpoint is yet to be uncovered. Moreover, casting the iterative approximation steps in the language of proximal operators on the space of density functions may provide theoretical insights and computational benefits. The purpose of the present paper has been to highlight and elucidate the ideas in [1] and [6] in the context of linear Gaussian systems. We hope that this study will help to motivate further exploration of this topic.

APPENDIX

In this Appendix, we collect some lemmas that are used in Sections III and IV. In addition, we will show in Corollary 1 below that applying Lemma 1 and 2 together enables us to provide an alternative proof of a Theorem in [11], which might be of independent interest.

Lemma 1: If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are symmetric positive definite matrices, then $\operatorname{tr}\left(\boldsymbol{X}^{\frac{1}{2}}\boldsymbol{Y}\boldsymbol{X}^{\frac{1}{2}}\right)^{\frac{1}{2}} \leq \sqrt{\operatorname{tr}(\boldsymbol{X})\ \operatorname{tr}(\boldsymbol{Y})}$.

Proof: From Uhlmann's variational formula (see [7], also Theorem 6.1 in [8]) , given any $\boldsymbol{G} \succ \boldsymbol{0}$, we have

$$\operatorname{tr}\left(\left(\boldsymbol{X}^{\frac{1}{2}}\boldsymbol{Y}\boldsymbol{X}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) \leq \sqrt{\operatorname{tr}(\boldsymbol{X}\boldsymbol{G})\ \operatorname{tr}(\boldsymbol{Y}\boldsymbol{G}^{-1})}, \tag{39}$$

where the equality in (39) is achieved for the specific choice $\boldsymbol{G}_{\text{opt}} = \boldsymbol{Y}^{\frac{1}{2}}\left(\boldsymbol{X}^{\frac{1}{2}}\boldsymbol{Y}\boldsymbol{X}^{\frac{1}{2}}\right)^{-\frac{1}{2}}\boldsymbol{X}^{\frac{1}{2}}\boldsymbol{Y}^{\frac{1}{2}}\boldsymbol{X}^{-\frac{1}{2}}$. Specializing (39) for $\boldsymbol{G}=\boldsymbol{Y}$, and noting that $\operatorname{tr}(\boldsymbol{X}\boldsymbol{Y}) \leq \operatorname{tr}(\boldsymbol{X})\operatorname{tr}(\boldsymbol{Y})$, the statement follows. ■

Lemma 2: Given a PDF $\varrho_0(\boldsymbol{x}) \in \mathscr{D}_2$ with mean $\boldsymbol{\mu}_0 \in \mathbb{R}^n$, and $n\times n$ covariance matrix $\boldsymbol{P}_0 \succ \boldsymbol{0}$. Then $\inf\limits_{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}} W_2^2(\varrho,\varrho_0)$ equals

$$\parallel \boldsymbol{\mu}-\boldsymbol{\mu}_0 \parallel_2^2 \ + \ \operatorname{tr}\left(\boldsymbol{P}+\boldsymbol{P}_0 - 2\left(\boldsymbol{P}_0^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{P}_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right), \tag{40}$$

and is achieved by push-forward of $\varrho_0(\boldsymbol{x})$ via an affine transport map $\boldsymbol{M}\boldsymbol{x}+\boldsymbol{m}$, where $\boldsymbol{M} := \boldsymbol{P}^{\frac{1}{2}}\left(\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{P}_0\boldsymbol{P}^{\frac{1}{2}}\right)^{-\frac{1}{2}}\boldsymbol{P}^{\frac{1}{2}}$, and $\boldsymbol{m} := \boldsymbol{\mu}-\boldsymbol{\mu}_0$, that is, the arginf for (40) is $\varrho(\boldsymbol{x}) = \sqrt{\frac{\det(\boldsymbol{P}_0)}{\det(\boldsymbol{P})}}\,\varrho_0\left(\boldsymbol{P}^{-\frac{1}{2}}\left(\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{P}_0\boldsymbol{P}^{\frac{1}{2}}\right)^{\frac{1}{2}}\boldsymbol{P}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})+\boldsymbol{\mu}_0\right)$. In particular, if $\varrho_0 = \mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{P}_0)$, then $\varrho = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{P})$.

Proof: Let ϱ_0 be as given, and choose any $\varrho \in \mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}$. Let $\overline{\varrho}_0$ and $\overline{\varrho}$ be obtained by translating ϱ_0 and ϱ respectively, such that both $\overline{\varrho}_0$ and $\overline{\varrho}$ have zero mean. Using (10), we can directly verify [9, p. 236] that $W_2^2(\varrho,\varrho_0) = \parallel \boldsymbol{\mu}-\boldsymbol{\mu}_0 \parallel_2^2 + W_2^2(\overline{\varrho},\overline{\varrho}_0)$. On the other hand, it is known [10, p. 11, Proposition 1.1.6] that

$$\begin{aligned} &W_2^2(\overline{\varrho},\overline{\varrho}_0) \geq \operatorname{tr}\left(\boldsymbol{P}+\boldsymbol{P}_0-2\left(\boldsymbol{P}_0^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{P}_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)\\ &\Rightarrow W_2^2(\varrho,\varrho_0) \geq \text{RHS of (40)}. \end{aligned} \tag{41}$$

Now consider a candidate transport map $\boldsymbol{M}\boldsymbol{x}+\boldsymbol{m}$ where $\boldsymbol{M}$ and $\boldsymbol{m}$ are functions of $\boldsymbol{P},\boldsymbol{P}_0,\boldsymbol{\mu},\boldsymbol{\mu}_0$ as in the statement. It suffices to prove that our candidate transport map indeed achieves the equality in (41). To this end, directly substituting the expressions for $\boldsymbol{M}$ and $\boldsymbol{m}$, notice that the push-forward has mean $\boldsymbol{M}\boldsymbol{\mu}_0+\boldsymbol{m}=\boldsymbol{\mu}$, and covariance $\boldsymbol{M}\boldsymbol{P}_0\boldsymbol{M}^\top = \boldsymbol{P}$. So our candidate transport map $(\boldsymbol{M},\boldsymbol{m})$ is feasible. To show optimality, from (10) notice that $W_2^2(\overline{\varrho},\overline{\varrho}_0) = \inf\limits_{\boldsymbol{C}\in\mathbb{R}^{d\times d}} \operatorname{tr}(\boldsymbol{P}+\boldsymbol{P}_0-2\boldsymbol{C})$, where $\boldsymbol{C} := \boldsymbol{M}\boldsymbol{P}_0$ solves $\boldsymbol{P}_0 - \boldsymbol{C}\boldsymbol{P}^{-1}\boldsymbol{C}^\top \succeq \boldsymbol{0}$, which has known optimal solution $\boldsymbol{C}_{\text{opt}} := \boldsymbol{M}_{\text{opt}}\boldsymbol{P}_0 = \boldsymbol{P}_0\boldsymbol{P}^{\frac{1}{2}}\left(\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{P}_0\boldsymbol{P}^{\frac{1}{2}}\right)^{-\frac{1}{2}}\boldsymbol{P}^{\frac{1}{2}}$. Since our candidate $\boldsymbol{M} := \boldsymbol{P}^{\frac{1}{2}}\left(\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{P}_0\boldsymbol{P}^{\frac{1}{2}}\right)^{-\frac{1}{2}}\boldsymbol{P}^{\frac{1}{2}}$ satisfies $\operatorname{tr}(\boldsymbol{M}\boldsymbol{P}_0) = \operatorname{tr}(\boldsymbol{M}_{\text{opt}}\boldsymbol{P}_0) = \operatorname{tr}\left(\left(\boldsymbol{P}_0^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{P}_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)$, the statement follows. ■

In the Corollary below, combining Lemma 1 and 2, we recover a result in [11, Theorem 3.1].

Corollary 1: Given PDF ϱ_0 with mean $\boldsymbol{\mu}_0$, covariance $\boldsymbol{P}_0 \succ \boldsymbol{0}$, suppose $\operatorname{tr}(\boldsymbol{P}_0)=\tau_0$. For fixed $\boldsymbol{\mu}$ and $\tau>0$,

$$\inf_{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\tau}} W_2^2(\varrho,\varrho_0) = \left(\sqrt{\tau}-\sqrt{\tau_0}\right)^2 \ + \ \parallel \boldsymbol{\mu}-\boldsymbol{\mu}_0 \parallel_2^2, \tag{42}$$

and is achieved by $\varrho(\boldsymbol{x}) = \left(\frac{\tau_0}{\tau}\right)^{\frac{d}{2}}\varrho_0\left(\frac{\tau_0}{\tau}(\boldsymbol{x}-\boldsymbol{\mu})+\boldsymbol{\mu}_0\right)$.

Proof: Let us choose $\boldsymbol{P} := \frac{\tau}{\tau_0}\boldsymbol{P}_0$, and from (40) observe that $\inf\limits_{\xi\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}} W_2^2(\xi,\varrho_0) = \left(\sqrt{\tau}-\sqrt{\tau_0}\right)^2 + \parallel \boldsymbol{\mu}-\boldsymbol{\mu}_0 \parallel_2^2$. On the other hand, for any $\varrho \in \mathscr{D}_{\boldsymbol{\mu},\tau}$, we know from (41) that

$$W_2^2(\varrho,\varrho_0) \geq \ \tau+\tau_0 - 2\ \operatorname{tr}\left(\boldsymbol{P}_0^{\frac{1}{2}}\boldsymbol{S}\boldsymbol{P}_0^{\frac{1}{2}}\right)^{\frac{1}{2}} \ + \ \parallel \boldsymbol{\mu}-\boldsymbol{\mu}_0 \parallel_2^2,$$

where $\boldsymbol{S}$ is the covariance of ϱ. Using Lemma 1, we get

$$\begin{aligned} &\operatorname{tr}\left(\boldsymbol{P}_0^{\frac{1}{2}}\boldsymbol{S}\boldsymbol{P}_0^{\frac{1}{2}}\right)^{\frac{1}{2}} \leq \sqrt{\tau\tau_0} \Rightarrow W_2^2(\varrho,\varrho_0) \geq \\ &\inf_{\xi\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}} W_2^2(\xi,\varrho_0) = \left(\sqrt{\tau}-\sqrt{\tau_0}\right)^2 \ + \ \parallel \boldsymbol{\mu}-\boldsymbol{\mu}_0 \parallel_2^2, \end{aligned} \tag{43}$$

and that the equality is achieved when $\boldsymbol{S}=\boldsymbol{P}=\frac{\tau}{\tau_0}\boldsymbol{P}_0$. In that case, $\det(\boldsymbol{S}) = \left(\frac{\tau}{\tau_0}\right)^d\det(\boldsymbol{P}_0)$, and hence Lemma 2 yields the arginf $\varrho(\boldsymbol{x})$ for (42) as

$$\begin{aligned} &\sqrt{\frac{\det(\boldsymbol{P}_0)}{\det(\boldsymbol{S})}}\,\varrho_0\left(\boldsymbol{S}^{-\frac{1}{2}}\left(\boldsymbol{S}^{\frac{1}{2}}\boldsymbol{P}_0\boldsymbol{S}^{\frac{1}{2}}\right)^{\frac{1}{2}}\boldsymbol{S}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})+\boldsymbol{\mu}_0\right)\\ &= \left(\frac{\tau_0}{\tau}\right)^{\frac{d}{2}}\varrho_0\left(\frac{\tau_0}{\tau}(\boldsymbol{x}-\boldsymbol{\mu})+\boldsymbol{\mu}_0\right). \end{aligned}$$

■

Lemma 3: If $\mathcal{E}(\cdot)$ depends on ϱ only via the mean and covariance of ϱ, then $\inf\limits_{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}} \mathcal{F}(\varrho)$ is achieved by $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{P})$.

Proof: As $\mathcal{E}(\varrho) \equiv \mathcal{E}(\boldsymbol{\mu},\boldsymbol{P})$, hence from (11) we get $\inf\limits_{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}} \mathcal{F}(\varrho) = \mathcal{E}(\boldsymbol{\mu},\boldsymbol{P}) + \beta^{-1}\inf\limits_{\varrho\in\mathscr{D}_{\boldsymbol{\mu},\boldsymbol{P}}} \int \varrho\log\varrho\, d\boldsymbol{x}$. Since $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{P})$ is the maximum entropy PDF under prescribed mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{P}$, hence the statement. ■

Lemma 4: For $\boldsymbol{P},\boldsymbol{P}_0 \succ 0$,

$$\frac{\partial}{\partial\boldsymbol{P}}\ \operatorname{tr}\left(\boldsymbol{P}_0^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{P}_0^{\frac{1}{2}}\right)^{\frac{1}{2}} = \frac{1}{2}\boldsymbol{P}_0^{\frac{1}{2}}\left(\boldsymbol{P}_0^{-\frac{1}{2}}\boldsymbol{P}^{-1}\boldsymbol{P}_0^{-\frac{1}{2}}\right)^{\frac{1}{2}}\boldsymbol{P}_0^{\frac{1}{2}}.$$

Proof: We refer the readers to Appendix B in [12]. ■

References

[1] 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.

[2] C. Villani, *Topics in Optimal Transportation*. Graduate Studies in Mathematics, Vol. 58, First ed., American Mathematical Society; 2003.

[3] J.-D. Benamou, and Y. Brenier, "A Computational Fluid Mechanics Solution to the Monge–Kantorovich Mass Transfer Problem". *Numerische Mathematik*. Vol. 84, No. 3, pp. 375–393, 2000.

[4] H. Risken, *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer Series in Synergetics, Vol. 18, First ed., Springer; 1989.

[5] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics, ETH Zürich, Second ed., Birkhäuser; 2008.

[6] R.S. Laugesen, P.G. Mehta, S.P. Meyn, and M. Raginsky, "Poisson's Equation in Nonlinear Filtering". *SIAM Journal on Control and Optimization*. Vol. 53, No. 1, pp. 501–525, 2015.

[7] A. Uhlmann, "The "Transition Probability" in the State Space of a *-Algebra". *Reports on Mathematical Physics*, Vol. 9, No. 2, pp. 273–279, 1976.

[8] D. Petz, *Quantum Information Theory and Quantum Statistics*. Theoretical and Mathematical Physics, First ed., Springer; 2008.

[9] C.R. Givens, and R.M. Shortt, "A Class of Wasserstein Metrics for Probability Distributions". *The Michigan Mathematical Journal*, Vol. 31, No. 2, pp. 231–240, 1984.

[10] S.T. Rachev, and L. Rüschendorf, *Mass Transportation Problems. Volume I: Theory*. First ed., Springer; 1998.

[11] E.A. Carlen, and W. Gangbo, "Constrained Steepest Descent in the 2-Wasserstein Metric". *Annals of Mathematics*, pp. 807–846, 2003.

[12] A. Halder, and E.D.B. Wendel, "Finite Horizon Linear Quadratic Gaussian Density Regulator with Wasserstein Terminal Cost". *Proceedings of the 2016 American Control Conference*, pp. 7249–7254, 2016.

[13] N.J. Higham, and H.M. Kim, "Solving A Quadratic Matrix Equation by Newton's Method with Exact Line Searches". *SIAM Journal on Matrix Analysis and Applications*, Vol. 23, No. 2, pp. 303–316, 2001.

[14] D. Liberzon, and R.W. Brockett, "Spectral Analysis of Fokker–Planck and Related Operators Arising from Linear Stochastic Differential Equations". *SIAM Journal on Control and Optimization*, Vol. 38, No. 5, pp. 1453–1467, 2000.

[15] R.W. Brockett, and J.C. Willems, "Stochastic Control and the Second Law of Thermodynamics". *Proceedings of the 1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, pp. 1007–1011, 1978.

[16] R.W. Brockett, "Notes on Stochastic Processes on Manifolds". *Systems and Control in the Twenty-first Century*, pp. 75–100, Springer; 1997.

[17] R.L. Stratonovich, "Application of the Theory of Markov Processes for Optimum Filtration of Signals". *Radio Eng. Electron. Phys. (USSR)*, Vol. 1, pp. 1–19, 1960.

[18] H.J. Kushner, "On the Differential Equations Satisfied by Conditional Densities of Markov Processes, with Applications". *Journal of the SIAM Series A Control*, Vol 2, No. 1, pp. 106–119, 1964.

[19] M. Fujisaki, G. Kallianpur, and H. Kunita, "Stochastic Differential Equations for the Non Linear Filtering Problem". *Osaka Journal of Mathematics*, Vol. 9, No. 1, pp. 19–40, 1972.

[20] R.E. Kalman, and R.S. Bucy, "New Results in Linear Filtering and Prediction Theory". *Journal of Basic Engineering*, Vol. 83, No. 3, pp. 95–108, 1961.

[21] A. Yezzi, and E.I. Verriest, "Nonlinear Observers via Regularized Dynamic Inversion". *Proceedings of the 2007 American Control Conference*, pp. 1693–1698, 2007.