Efficient Bayesian Inference Methods via Convex Optimization and Optimal Transport

Sanggyun Kim, Rui Ma, Diego Mesa, and Todd P. Coleman

***Abstract*—In this paper, we consider many problems in Bayesian inference - from drawing samples to posteriors, to calculating confidence intervals, to implementing posterior matching algorithms, by finding maps that push one distribution to another. We show that for a large class of problems (with log-concave likelihoods and log-concave priors), these problems can be efficiently solved using convex optimization. We provide example applications within the context of dynamic statistical signal processing.**

I. Introduction

Imagine that we have X that is drawn a priori according to $P_X \in \mathsf{P}(\mathsf{X})$, where $\mathsf{X} \subset \mathbb{R}^d$. We assume that P_X has a density with respect to Lebesgue measure, given by $f_X(x)$. We also assume that there is a likelihood function $f_{Y|X=x}$. We observe $Y = y$. As such, the posterior distribution $P_{X|Y=y}$ has a density given by $f_{X|Y=y}(x)$, determined by Bayes' rule:

$$f_{X|Y=y}(x) = \frac{f_{Y|X}(y|x) f_X(x)}{\beta_y} \tag{1}$$

where β_y is a constant that does not vary with x, given by:

$$\beta_y = \int_{v \in \mathsf{X}} f_{Y|X}(y|v) f_X(v) dv \tag{2}$$

In general, calculation of β_y is hard.

In many cases, it is desirable to construct the posterior expectation $E[X|Y=y]$. Typical approaches to perform this involve Monte Carlo methods [1]. Markov Chain Monte Carlo (MCMC) methods, where samples are drawn from a Markov chain whose invariant distribution is that of the posterior, are particularly common. One problem, however, with these methods is that because they are drawn from a Markov chain, they are necessarily statistically **dependent**; so the law of averages kicks in more slowly. Developing an alternative approach to draw independent, identically distributed (i.i.d.) samples $Z_1, \ldots, Z_n$ from $P_{X|Y=y}$ to estimate $E[X|Y=y]$ as

$$\hat{E}[X|Y=y] = \frac{1}{N} \sum_{i=1}^{N} Z_i$$

would be particularly attractive. In filtering problems, X and Y are random processes $(X_i, Y_i : i \geq 1)$ and so it is desirable to constrict a *filter* that constructs $P_{X_i|Y^i=y^i}$ from $P_{X_{i-1}|Y^{i-1}=y^{i-1}}$ and the most recent observation Y_i. In these settings, it is again crucially important to have β_y to perform these computations. Here, we will develop an alternative approach that is inspired from the theory of optimal transportation. We show that for a large class of problems (with log-concave likelihoods and log-concave priors), these problems can be efficiently solved using convex optimization. We provide example applications within the context of dynamic statistical signal processing.

II. Definitions

Consider a set $\mathsf{W} \subset \mathbb{R}^d$ for some d. Define the space of all probability measures on W as $\mathsf{P}(\mathsf{W})$.

Definition II.1 (Push-forward)**.** *Given a $P \in \mathsf{P}(\mathsf{W})$ and a $Q \in \mathsf{P}(\mathsf{W})$, we say that a map $S : \mathsf{W} \to \mathsf{W}$ pushes P to Q (denoted as $S_{\#}P = Q$) if a random variable W with distribution P results in $Z \triangleq S(W)$ having distribution Q.*

We say that $S : \mathsf{W} \to \mathsf{W}$ is a 'diffeomorphism' on W if S is invertible and both S and S^{-1} are differentiable. Denote the set of all diffeomorphisms on W as $\mathcal{S}(\mathsf{W})$. With this, we have the following lemma from standard probability:

Lemma II.2. *Consider a diffeomorphism $S \in \mathcal{S}(\mathsf{W})$ and $P, Q \in \mathsf{P}(\mathsf{W})$ that both have densities p, q with respect to the Lebesgue measure. Then $S_{\#}P = Q$ if and only if*

$$p(u) = q(S(u)) \left|\det\left(J_S(u)\right)\right| \quad \text{for all } u \in \mathsf{W} \tag{3}$$

where $\left|\det\left(J_S(u)\right)\right|$ is the absolute value of the determinant of the Jacobian of S at u.

III. Optimal Transport and Bayesian Inference

We note now that if we were able to find a diffeomorphism S_y^* for which $S_{y\#}^* P_X = P_{X|Y=y}$, then we could associate p with f_X and q with $f_{X|Y=y}$ in (3) to obtain:

$$\begin{aligned} f_X(x) &= f_{X|Y=y}(S_y^*(x)) \left|\det\left(J_{S_y^*}(x)\right)\right| \\ &= \frac{f_{Y|X}(y|S_y^*(x)) f_X\left(S_y^*(x)\right)}{\beta_y} \left|\det\left(J_{S_y^*}(x)\right)\right| \end{aligned} \tag{4}$$

where (4) comes from Bayes' rule (1). We now define the operator T as

$$\begin{aligned} T(S,x) &\triangleq \log f_{Y|X}(y|S(x)) + \log f_X\left(S(x)\right) \\ &+ \log\left|\det\left(J_S(x)\right)\right| - \log f_X(x) \end{aligned} \tag{5}$$

and thus we note that (4) can be equivalently stated as

Lemma III.1. *A diffeomorphism S_y^* satisfies $S_{y\#}^* P_X = P_{X|Y=y}$ if and only if*

$$T(S_y^*, x) \equiv \log \beta_y, \quad \text{for all } x \in \mathsf{X}. \tag{6}$$

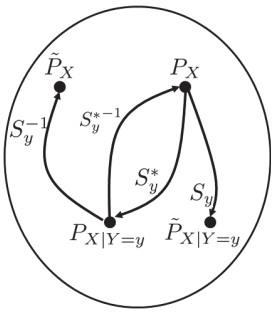

Fig. 1. We begin with a prior distribution P_X. Upon an observation $Y = y$, it is our objective to find $P_{X|Y=y}$. Up front, because of the difficulty in computing β_y, $P_{X|Y=y}$ is unknown; but we know it exists. A 'desirable' diffeomorphism S_y^* pushes the prior P_X to the posterior $P_{X|Y=y}$; equivalently, ${S_y^*}^{-1}$ pushes the posterior $P_{X|Y=y}$ to the prior P_X. An arbitrary diffeomorphism S_y will push P_X to some distribution $\tilde{P}_{X|Y=y}$ that is not necessarily $P_{X|Y=y}$; equivalently, ${S_y}^{-1}$ pushes the posterior $P_{X|Y=y}$ to some distribution $\tilde{P}_X$ that is not necessarily P_X.

Note that this encodes a variational principle. The LHS of the above equation is allowed to vary with x. But for S_y^*, at **any** x, $T(S_y^*, x)$ takes on the same value.

This suggests one particular problem formulation, problem $P1$ [2]:

$$\begin{aligned}(P1) \qquad S_y^* &= \underset{S_y \in \mathcal{S}(\mathsf{X})}{\arg\min}\, V_1(S_y), \\ V_1(S_y) &\triangleq \int_{x\in\mathsf{X}} \left|T(S_y,x) - \mathbb{E}[T(S_y,X)]\right|^2 f_X(x)dx\end{aligned}$$

Remark 1. *As we will see, attempting to solve problem P1 is hard. Even if we try trying to find a diffeomorphism S_y^* for which ${S_y^*}_{\#}P_X = P_{X|Y=y}$ using a linear combination of basis functions (e.g. polynomial chaos expansion) by solving problem P1, we still have a non-convex problem, even for many natural priors and likelihoods. Our insight is to abandon the approach of problem P1 espoused in [2] and instead focus on a subset of common problems with log-concave structure, using an alternative KL divergence based criterion. This leads to an efficient algorithms via convex optimization.*

We now show that we can find a diffeomorphism S_y^* for which ${S_y^*}_{\#}P_X = P_{X|Y=y}$ using an an alternative optimality criterion that results in efficient computational algorithms using linear basis expansions for many natural priors and likelihoods. Consider any other diffeomorphism S_y. See Figure 1. Note that clearly ${S_y^{-1}}_{\#}P_{X|Y=y} = \tilde{P}_X$ for some $\tilde{P}_X$. Denote its density as $\tilde{f}_X(x)$. From the Jacobian equation (3):

$$\begin{aligned}\tilde{f}_X(x) &= f_{X|Y=y}(S_y(x))\left|\det\left(J_{S_y}(x)\right)\right| \\ &= \frac{f_{Y|X}(y|S_y(x))f_X\left(S_y(x)\right)}{\beta_y}\left|\det\left(J_{S_y}(x)\right)\right| \quad (7)\end{aligned}$$

Note that by careful inspection of (7) and (5), we have that

$$\log\frac{f_X(x)}{\tilde{f}_X(x)} = \log\beta_y - T(S_y,x). \quad (8)$$

So if a diffeomorphism S_y^* satisfies ${S_y^*}_{\#}P_X = P_{X|Y=y}$, then $f_X(x) = \tilde{f}_X(x)$ and so:

$$0 = \log\beta_y - T(S_y^*,x) \Leftrightarrow T(S_y^*,x) = \log\beta_y, \quad \text{for all } x. \quad (9)$$

For any diffeomorphism S_y, we have:

$$\begin{aligned}D\Big(P_X\|\tilde{P}_X\Big) &\triangleq \int_{x\in\mathsf{X}} f_X(x)\log\frac{f_X(x)}{\tilde{f}_X(x)}dx \\ &= \log\beta_y - \int_{x\in\mathsf{X}} f_X(x)T(S_y,x)dx.\end{aligned}$$

We now state one useful property of the KL divergence:

Lemma III.2. *The KL divergence satisfies $D\Big(P_X\|\tilde{P}_X\Big) \geq 0$ and is 0 if and only if $P_X = \tilde{P}_X$.*

So another way to find a diffeomorphism S_y^* for which ${S_y^*}_{\#}P_X = P_{X|Y=y}$ is to solve:

$$\begin{aligned}(P2) \qquad S_y^* &= \underset{S_y\in\mathcal{S}(\mathsf{X})}{\arg\max}\, V_2(S_y), \\ V_2(S_y) &\triangleq \int_{x\in\mathsf{X}} f_X(x)T(S_y,x)dx. \quad (10)\end{aligned}$$

Lemma III.3. *Problem P1 and P2 have the same optimal solutions. Moreover, for any diffeomorphism S_y^*, ${S_y^*}_{\#}P_X = P_{X|Y=y}$ holds if and only if it is an optimal solution to problem P1 (P2).*

Taking a weighted sum of the T operator is appealing, except for the issue of trying to taze out maps for which $\det\left(J_{S_y}(x)\right) > 0$, and others for which $\det\left(J_{S_y}(x)\right) < 0$, - where $J_{S_y}(x) \equiv S_y'(x)$. Note that both classes of maps are equally as good; however, from an optimization viewpoint, this leads to non-convexity in an optimization problem P2. As such, we plan to try to find only one such solution satisfying the Jacobian equation, for which it can be efficiently found. We instead consider finding maps for which $\det(S_y(x)) > 0$ and so we consider the following problem P3:

$$\begin{aligned}(P3) \qquad S_y^* &= \underset{S_y\in\mathcal{S}(\mathsf{X}),\ J_{S_y}(x)\succ 0\ \forall x\in\mathsf{X}}{\arg\max} V_3(S_y), \\ V_3(S_y) &\triangleq \int_{x\in\mathsf{X}} f_X(x)\tilde{T}(S_y,x)dx \quad (11) \\ \tilde{T}(S,x) &\triangleq \log f_{Y|X}(y|S(x)) + \log f_X\left(S(x)\right) \\ &+ \log\det\left(J_S(x)\right) - \log f_X(x) \quad (12)\end{aligned}$$

Note that the two key differences in problem P2 and P3 are:

- the feasible set in P3 is constrained so that J_{S_y} is positive definite
- In P2, $T(S,x)$ has a $\log|\det\left(J_S(x)\right)|$ term whereas in P3, $\tilde{T}(S,x)$ has a $\log\det\left(J_S(x)\right)$ term

We can now state this crucially important problem:

Theorem III.4. *There exists an optimal solution S_y^* to Problem P3 and it is also an optimal solution to problem P2.*

Proof: Note that the feasible set in problem P3 is clearly a subset of the feasible set in problem P2. Moreover, for any

positive definite matrix A, $\det(A) > 0$. Therefore $\tilde{T}(S,x) = T(S,x)$ for any S in the feasible set of problem $P3$. Therefore, clearly $V_2^* \geq V_3^*$. We relate this to the Monge-Kontarovich problem with Euclidean distance cost [3]:

$$(MK) \quad \min_{S_y:\mathsf{X}\to\mathsf{X}} \quad \int_{x\in\mathsf{X}} f_X(x)\|x - S_y(x)\|^2 dx \quad s.t. \quad {S_y}_{\#}P_X = P_{X|Y=y}$$

Some key properties of problem MK are as follows [3]: (i) S_y^* is a diffeomorphism (i.e. $S_y^* \in \mathcal{S}(\mathsf{X})$) and (ii) $S_y^*(x) = \nabla h(x)$ where h is a strictly convex function. As such, $J_{S_y^*}(x) = J_{\nabla h}(x) \succ 0$ for any x: $a^T J_{S_y^*}(x) a > 0$ for all a and x. This implies that there exists an S_y^* in the feasible set of problem P3 for which ${S_y^*}_{\#}P_X = P_{X|Y=y}$. It thus directly follows that $V_2^* = V_2(S_y^*) = V_3^* = V_3(S_y^*)$. ∎

IV. A Provably Good and Computationally Efficient Method to Calculate the Posterior

Note that problem P3 is a search over a space of functions, and is thus in general not computationally feasible. However, using optimal transport, we have been able to constrain the search to be over monotonic functions S_y and still find a solution to the Jacobian Equation. This observation, along with appropriate assumptions (e.g. log-concavity) on the prior and likelihood, lead to a computationally efficient method via convex optimization.

A. Linear Basis Expansion

Rather than optimizing over functions, we can perform functional analysis and approximate any $S \in \mathcal{S}(\mathsf{X})$ as a linear combination of basis functions

$$S(x) = \sum_{j\in\mathcal{J}} g_j \phi^{(j)}(x), \tag{13}$$

where $\phi^{(j)}(x) \in \mathbb{R}$ and $g_j \in \mathbb{R}^d$, with d being the dimension of W. One natural way to do this, for example if $\mathsf{X} \subset \mathbb{R}$, is to perform a polynomial chaos expansion, where

$$\int_{x\in\mathsf{X}} \phi^{(i)}(x)\phi^{(j)}(x) f_X(x) dx = \begin{cases} C_i, & i = j \\ 0, & i \neq j \end{cases}.$$

For example, if $\mathsf{X} = [-1,1]$ and P_X is uniformly distributed, then $\phi^{(j)}(x)$ are the Legendre polynomials. If $\mathsf{X} = \mathbb{R}$ and P_X is Gaussian, then $\phi^{(j)}(x)$ are the Hermite polynomials.

Remark 2. *In principle, any basis of polynomials for which the truncated expansion of functions is dense in the space of all functions on X suffices. Using the polynomial chaos expansion where orthogonality is measured with respect to the prior, means that computing conditional expectations and other calculations can be done only with linear algebra.*

Now define $K = |\mathcal{J}|$ and we have that for $\mathsf{X} \subset \mathbb{R}$:

$$\begin{aligned} F &= [g_1, \ldots, g_K], && d \times K && (14)\\ A(x) &= [\phi^{(1)}(x), \ldots, \phi^{(K)}(x)]^T, && , K \times 1 && (15)\\ S(x) &= FA(x), && d \times 1 && (16)\\ J_A(x) &= \left[\frac{\partial \phi^{(i)}}{\partial x_j}(x)\right]_{i,j} && K \times d && (17)\\ J_S(x) &= FJ_A(x) && d \times d. && (18) \end{aligned}$$

Note that as $K \to \infty$, we develop a richer and richer set of candidate functions representative of $\mathcal{S}(\mathsf{X})$. Extension to $\mathsf{X} \subset \mathbb{R}^d$ can be done using tensor products, where

$$\phi^{(j)}(x) = \prod_{a=1}^{d} \psi_{j_a}(x_a).$$

and $j \to (j_1, \ldots, j_d)$ is defined in a standard diagonal manner. For example, if $d = 2$, then we have $j \to (j_1, j_2)$ is constructed as:

$$\begin{array}{lll} 0 \to (0,0) & 1 \to (0,1) & 3 \to (0,2) \\ 2 \to (1,0) & 4 \to (1,1) & \ldots \\ 5 \to (2,0) & \ldots & \end{array}$$

and so

$$\begin{aligned} \phi^{(0)}(x) &= \psi_0(x_1)\psi_0(x_2) && (19)\\ \phi^{(1)}(x) &= \psi_0(x_1)\psi_1(x_2) && (20)\\ \phi^{(2)}(x) &= \psi_1(x_1)\psi_0(x_2) && (21)\\ \phi^{(3)}(x) &= \psi_0(x_1)\psi_2(x_2) && (22)\\ \ldots &\quad \ldots && (23) \end{aligned}$$

B. A Convex Optimization Problem

With this linear basis expansion, we now show that under appropriate assumptions on the prior and likelihood, the T operator becomes log-concave. Under the basis expansion in (13), we have that $\log\det\left(J_{S_y}(x)\right) \equiv \log\det\left(FJ_A(x)\right)$ and so we re-define $T(S,x)$ in (5) as:

$$\begin{aligned} \tilde{T}(F,x) &\triangleq \log f_{Y|X}(y|FA(x)) + \log f_X\left(FA(x)\right) \\ &+ \log\det\left(FJ_A(x)\right) - \log f_X(x). \end{aligned} \tag{24}$$

We now define problem P4, that is in essence the same problem as P3 but where we approximate an expectation by a weighted sum of IID samples, and we approximate the set of all functions using a truncated polynomial chaos expansion:

$$(P4) \quad F^* = \underset{F\in\mathbb{R}^{d\times K}: FJ_A(X_1)\succ 0,\ldots,FJ_A(X_N)\succ 0}{\arg\max} V_4(F),$$

$$V_4(F) \triangleq \frac{1}{N}\sum_{i=1}^{N} \tilde{T}(F, X_i) \tag{25}$$

where $X_1, X_2, \ldots, X_N$ are drawn i.i.d. from P_X.

Lemma IV.1. *If $f_X(x)$ is log-concave and $f_{Y|X}(y|x)$ is log-concave in x, then $f_{X|Y=y}(x)$ is log-concave and problem P4 is a convex optimization problem.*

Proof: That $f_{X|Y=y}(x)$ is log-concave in x is trivial: it follows directly from the assumption that $f_X(x)$ and

$f_{X|Y=y}(x)$ are log-concave in x, along with Bayes' rule (1). As for showing that $\tilde{T}(F,x)$ is concave, this follows from (i) the assumption that $f_X(x)$ and $f_{X|Y=y}(x)$ are log-concave in x; and (ii) that concavity is preserved under affine transformations ($F \to FA(x)$, $F \to FJ_A(x)$) [4]. As for the feasible set, a set of vectors satisfying an affine positive definite constraint is convex [4]. ■

V. Applications of this Framework

A. Bayesian Confidence Intervals

In many parameter estimation situations, we have a likelihood function $P_{Y|W=w}$ and a prior P_W on the parameters. It is a common desire to efficiently find a confidence interval, i.e. an interval $I \subset \mathsf{W}$ s.t. $\mathbb{P}(W \in I|Y=y) \geq 0.95$. Typical ways to perform this involve Monte Carlo simulation [5]. We can solve this problem by first picking an interval $\tilde{I}$ s.t. $\mathbb{P}\left(W \in \tilde{I}\right) \geq 0.95$, which is trivial for common (e.g. Gaussian or uniform) priors, and then applying $I = S_y(\tilde{I})$ where $S^*_{y\,\#}P_W = P_{W|Y=y}$.

B. Sampling from Log-Concave Priors and Posteriors

In many cases, it is important to sample from a log-concave prior or a log-concave posterior. We first consider the scenario of sampling form a log-concave prior. Here, if we assume that $\mathsf{X} \subset \mathbb{R}^d$ is compact, then we note that we can imagine that $f_X(x)$ which is log-concave can be interpreted as

$$f_X(x) \equiv f_{\tilde{Y}|\tilde{X}=x}(y)$$

and $f_{\tilde{X}}(x) = C$. As such, then all we want to do is develop a method to construct a map that will map a prior distribution $f_{\tilde{X}}(x)$ that is uniform on X to a posterior distribution $f_{\tilde{X}|Y=y}$. Then, applying our framework will solve that problem. Moreover, we can compute whatever means we like by using the polynomial chaos expansion.

C. Dynamic Bayesian Analysis of Neural Data

It is quite common in dynamic anlaysis of neural data that Bayesian methods are used. In most cases, the likelihood and prior are log-concave. In many situations, to model plasticity or to model motor kinematics in a brain-machine interface, or when modeling with place cells, a state-space Markov model is assumed on the latent process. In most cases, it has a Gaussian state space model, which will equate to log-concavity being maintained in the posterior $P_{X_t|Y^t=y^t}$. Most techniques are heuristic and make Gaussian approximations or use Markov chain Monte Carlo to recursively compute the posterior [6]–[9].

D. Time-Varying Causality Measures

It is quite common that one would like to understand non-stationarities in data to perform causal inference [10]. Typically, we define $p_t \triangleq P_{Y_t|Y^{t-1}=y^{t-1}}$ and $\tilde{p}_t \triangleq P_{Y_t|Y^{t-1},X^t=y^{t-1},x^t}$ and compute the directed information to develop a quantitative measure of causality:

$$I(X \to Y) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[D\left(\tilde{P}_t \| P_t\right)\right] \tag{26}$$

where the outer expectation is taken with respect to P_{Y^{t-1},X^t}. However, typical estimation methods for this assume stationarity and ergodicity of the joint processes to apply laws of large number methods to estimate and use laws of large numbers to estimate (26) for large T [10]–[12]. Here, we address computing a time-varying causality measure in spirit with Granger's philosophical notion of causality [13]. Imagine we first focus on computing p_t. In order to track possibly non-stationary changes, one way to do this is to use a regret minimization approach and implement a Bayesian mixture of experts approach [14]. For any $u \in \mathsf{E}$, given y^{t-1}, expert u gives a suggestion to specify $p_t^u \equiv P_{Y_t|Y^{t-1}=y^{t-1},\theta=u}$. A provably good and natural mixture strategy is to build p_t as a weighted combination of each of the p_t^u for every $u \in \mathsf{E}$:

$$p_t(y) = \int_{u\in\mathsf{E}} w_t(u) p_t^u(y) du \tag{27}$$

where w_t is non-negative and integrates to 1. $w_t(u)$ is exactly the **posterior distribution** on θ, with the prior being the initial weighting function:

$$w_t(u) = \frac{\overbrace{f_\theta(u)}^{\text{prior}}\overbrace{P_{Y^{t-1}|\theta=u}(y^{t-1})}^{\text{likelihood}}}{\beta_{y^{t-1}}} \tag{28}$$

$$f_\theta(u) \equiv w_0(u). \tag{29}$$

For many cases (i.e. exponential families), the log-likelihood is log-concave and thus so is $P_{Y^{t-1}|\theta=u}(y^{t-1}) = \prod_{k=1}^{t-1} p_k^u(y_k)$. As such, when $f_\theta(u)$ is log-concave, $w_t(u)$ can be calculated efficiently and recursively. Thus we can combine $w_t(u)$ and p_t^u for each u to construct our mixture predictor p_t by implementing (27) as follows:

$$P_{Y_t|Y^{t-1}=y^{t-1}}(a) = \frac{P_{\theta|Y^{t-1}=y^{t-1}}(u) P_{Y_t|Y^{t-1}=y^{t-1},\theta=u}(a)}{P_{\theta|Y^{t-1}=y^{t-1},Y_t=a}(u)}$$

$$\Leftrightarrow p_t(a) = \frac{w_t(u)p_t^u(a)}{\underbrace{P_{\theta|Y^{t-1}=y^{t-1},Y_t=a}(u)}_{w_{t+1}(u) \text{ with } Y_t=a}} \tag{30}$$

$\tilde{p}_t$ will be computed analogously to (30) using a standard likelihood model of $\tilde{p}_t^u \triangleq P_{Y_t|Y^{t-1}=y^{t-1},X^t=x^t,\theta=u}$:

$$\tilde{p}_t(a) = \frac{\tilde{w}_t(u)\tilde{p}_t^u(a)}{\underbrace{P_{\theta|Y^{t-1}=y^{t-1},Y_t=a,X^{t-1}=x^{t-1}}(u)}_{\tilde{w}_{t+1}(u) \text{ with } Y_t=a}} \tag{31}$$

As such, p_t and $\tilde{p}_t$ can be recursively calculated, to obtain the time-varying causal metric:

$$C_{X\to Y}(t) = D(\tilde{p}_t \| p_t)\,. \tag{32}$$

Figure 2 gives an example where neural spike trains simultaneously recorded in area M1 of a monkey were analyzed to understand how the dynamics of interacting neural processes relate to an external cue (that provides a new target to reach for). In the top, three time intervals were pre-selected (before, during, after cue) and static directed information was estimated using [11]. In the bottom, the time-varying method of (32)

was used for a point process GLM likelihood model [11] with Gaussian prior. More details of dynamics are illustrated in the bottom. For example, the directed relationship between neurons $8 \to 1$, $10 \to 1$, $6 \to 7$, and $1 \to 3$ in the bottom reflect the presence/absence of arrows in the top, yet the bottom uncovers more details of the 'on-off' causal dynamics.

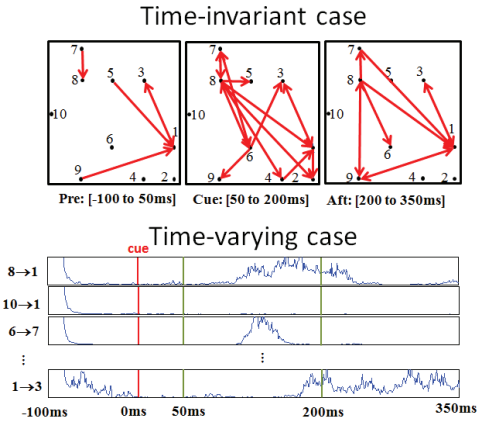

Fig. 2. Top: estimates of directed information between neural activity using (26). A directed arrow from neuron X to Y is present if $\mathbb{P}\left(I(X \to Y) > 0\right) > \epsilon$. Bottom: time-varying measures of causal relationships between neural activity using (32).

E. Minimax Optimal Regret in Sequential Prediction

For sequential prediction with respect to experts in E, again define expert u as p_t^u. Define the prediction made as p_t which is again a function of y^{t-1}. Under the log loss $l(p_t, y_t) \triangleq -\log p_t(y_t)$, the regret for predictor p with respect to E is

$$R_T(p, \mathsf{E}) \triangleq \sup_{y^T} \sum_{t=1}^{t} l(p_t, y_y) - \inf_{u \in \mathsf{E}} \sum_{t'=1}^{T} l(p_t^u, y_t). \qquad (33)$$

The best attainable regret, known as the minimax regret, is $R_T(\mathsf{E}) = \min_p R_T(p, \mathsf{E})$. The distribution p^* that attains $R_T(\mathsf{E})$ is the normalized maximum likelihood distribution, but it is undesirable because it is intractable to compute and does not have the prequential property [15]. Under very general conditions, a Baysian predictor (27) with Jeffreys' prior [16] asymptotically attain $R_T(\mathsf{E}) + o(T)$ [15]. As discussed in [17], many exponential family likelihoods (which are log-concave) have a log-concave Jeffrey's prior; in those settings, our efficient approach is applicable.

F. Posterior Matching for Feedback Communication

In "message-point" feedback communication problems (e.g. brain-machine interfaces [18]), the message lies in a continuum $\mathsf{W} \subset \mathbb{R}^d$ with a prior density satisfying $f_W(u) = C$ over W. Then, the channel input is $X_1 = \phi(W)$ and a noisy channel maps X_1 to $Y_1 = y$. With this, we obtain a posterior $P_{W_1|Y_1=y} \equiv P_{W|Y=y}$. With feedback, we would like to construct an $S_y^* \in \mathcal{S}(\mathsf{W})$ for which $W_2 = S_y(W_1)$ satisfies $P_{W_2|Y_1=y} \equiv P_W$ and $X_2 = \phi(W_2)$. This *posterior matching* [19] principle satisfies the necessary and sufficient conditions to maixmize mutual information $I(W; Y_1, Y_2)$. For time-invariant memoryless channels, we can use the same set of functions $\{S_y : y \in \mathsf{Y}\}$ for any t: implementing $\left(W_t = S_{y_{t-1}}(W_{t-1}) : t \geq 2\right)$ and $X_t = \phi(W_t)$ results in $I(W; Y^t) = tI$ [19]. For $\mathsf{W} = [0, 1]$, this was done in [19]. But for $\dim(\mathsf{W}) > 1$, we can apply an optimal transport framework to find such a scheme [20]. Here, we can develop an *efficient* method using the convex optimization framework that is dual to the Bayesian inference problem: obtain a map $S_{y,\text{PM}}^* \in \mathcal{S}(\mathsf{W})$ for which $S_{y,\text{PM}\#}^* P_{W|Y=y} = P_W$. We can equivalently find the inverse of S_y^* that pushes P_W to $P_{W|Y=y}$.

References

[1] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. Springer, 2008.

[2] T. El Moselhy and Y. Marzouk, "Bayesian inference with optimal maps," *Journal of Computational Physics*, 2012.

[3] C. Villani, *Topics in optimal transportation*. AMS, 2003.

[4] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[5] L. Eberly and G. Casella, "Estimating Bayesian credible intervals," *Journal of statistical planning and inference*, 2003.

[6] A. Smith and E. Brown, "Estimating a state-space model from point process observations," *Neural Computation*, 2003.

[7] U. Eden, L. Frank, R. Barbieri, V. Solo, and E. Brown, "Dynamic analysis of neural encoding by point process adaptive filtering," *Neural Computation*, vol. 16, no. 5, pp. 971–998, 2004.

[8] Z. Chen, F. Kloosterman, M. Wilson, and E. Brown, "Variational Bayesian inference for point process generalized linear models in neural spike trains analysis," in *IEEE ICASSP*, 2010.

[9] K. Yuan, M. Girolami, and M. Niranjan, "Markov Chain Monte Carlo Methods for State-Space Models with Point Process Observations," *Neural Computation*, pp. 1–25, 2012.

[10] C. Quinn, T. Coleman, N. Kiyavash, and N. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, 2011.

[11] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown, "A Granger causality measure for point process models of ensemble neural spiking activity," *PLoS Comput Biol*, vol. 7, no. 3, March 2011.

[12] J. Jiao, H. Permuter, L. Zhao, Y. Kim, and T. Weissman, "Universal estimation of directed information via sequential probability assignments," in *IEEE ISIT*, 2012.

[13] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 1969.

[14] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.

[15] P. Grünwald, *The minimum description length principle*. The MIT Press, 2007.

[16] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, 2002.

[17] M. R. M. D. Kim, S. and T. Coleman, "A class of efficiently constructable minimax optimal sequential predictors," in *Allerton*, 2012.

[18] C. Omar, A. Akce, M. Johnson, T. Bretl, R. Ma, E. Maclin, M. McCormick, and T. Coleman, "A feedback information-theoretic approach to the design of brain–computer interfaces," *Intl. Journal of Human–Computer Interaction*, vol. 27, no. 1, pp. 5–23, 2010.

[19] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Trans. Inf. Theory*, 2011.

[20] R. Ma and T. Coleman, "Generalizing the posterior matching scheme to higher dimensions via optimal transportation," in *Allerton*, 2011.