ISIT'2013 1569717923

Minimax Filtering Regret via Relations Between Information and Estimation

Albert No
Stanford University
Email: albertno@stanford.edu

Tsachy Weissman
Stanford University
Email: tsachy@stanford.edu

***Abstract*—We investigate the problem of continuous-time causal estimation under a minimax criterion. Let $X^T = \{X_t, 0 \leq t \leq T\}$ be governed by probability law P_θ from some class of possible laws indexed by $\theta \in \mathcal{S}$, and Y^T be the noise corrupted observations of X^T available to the estimator. We characterize the estimator minimizing the worst case regret, where regret is the difference between the expected loss of the estimator and that optimized for the true law of X^T.**

We then relate this minimax regret to the channel capacity when the channel is either Gaussian or Poisson. In this case, we characterize the minimax regret and the minimax estimator more explicitly. If we assume that the uncertainty set consists of deterministic signals, the worst case regret is exactly equal to the corresponding channel capacity, namely the maximal mutual information attainable across the channel among all possible distributions on the uncertainty set of signals. Also, the optimum minimax estimator is the Bayesian estimator assuming the capacity-achieving prior. Moreover, we show that this minimax estimator is not only minimizing the worst case regret but also essentially minimizing the regret for "most" of the other sources in the uncertainty set.

We present a couple of examples for the construction of an approximately minimax filter via an approximation of the associated capacity achieving distribution.

I. Introduction

Recent relations between information and estimation have shown fundamental links between the causal estimation error and information theoretic quantities. In [1], Duncan showed that causal estimation error of an additive white Gaussian noise (AWGN) corrupted signal is equal to the mutual information between the input and output processes divided by signal-to-noise ratio. In [2], Weissman extended the result to the scenario of mismatched estimation, where the estimator assumes that the input signal is governed by a law Q while its true law is P. In this case, the cost of mismatch, which is half the difference between the mismatched causal estimation error and the optimal (non-mismatched) causal estimation error, is given by the relative entropy between the laws of output processes when the input processes have laws P and Q, respectively. In [3], Atar et. al. showed that similar information-estimation relations exist in the Poisson channel for both mismatched and non-mismatched settings.

In this paper, we investigate the continuous-time causal estimation problem. We assume that the input process is governed by a probability law from a known uncertainty class $\mathcal{P}$ where the estimator does not know the true law. In particular, suppose that the input process is governed by law $P_\theta \in \mathcal{P}$, where $\theta \in \mathcal{S}$ and $\mathcal{S}$ is the uncertainty set known to decoder. In this setting, it is natural to consider the minimax estimator which minimizes the worst case regret, where regret is defined as the difference between the causal estimation error of the estimator and that of the optimal estimator. One of the main contributions of this paper is characterizing the minimax estimator. We show that it is in fact a Bayesian estimator under a distribution which is the capacity-achieving mixture of distributions associated with the channel whose input is a source in the uncertainty set.

We can find similar arguments in the classical universal source coding theory. Redundancy capacity theory in this setting tells us that the minimax redundancy, which is the minimum of the worst case redundancy, coincides with the maximum mutual information between input and output of a channel whose input is a choice of a law from the uncertainty set and whose output is a realization of that law. We can combine the results of mismatched estimation and the above redundancy capacity theorem in order to relate the minimax regret to the corresponding mutual information, if the channel is either Gaussian or Poisson. Indeed, the corresponding minimax regret turns out to be equal to the mutual information between the input index and the corresponding output which we shall refer to as "regret capacity". Moreover, the optimal minimax filter is Bayesian with respect to the same prior that achieves maximum mutual information. Therefore, if we know the distribution that maximizes mutual information, we can induce the optimal minimax estimator. Further, we shall see that if the class of measures $\mathcal{P}$ is a set of deterministic signals, this mutual information simplifies to the mutual information between input and output processes X^T and Y^T. This allows us to harness well known results from channel coding to characterize and construct the optimum minimax filter.

Since, by definition, the goal in minimax estimation is to minimize the worst case estimation regret, one possible critique is that it might not result in good estimation for many of sources in the class. However, in universal source coding theory, Merhav and Feder [4] showed that the minimax encoder works well for "most" distributions in the uncertainty set, where "most" is measured with respect to the capacity-achieving prior which is argued to be the "right" prior. Indeed, the framework of [4] strengthened and generalized results of this nature that were established for parametric uncertainty sets by Rissanen in [5]. We can apply this idea to our mini-

max estimation setting. These results imply that the minimax estimator not only minimizes the worst case error, but does essentially as well as the optimal estimator for most sources.

Our results for the Gaussian and the Poisson channel carry over to accommodate the presence of feedback. We show that they are still valid in the presence of feedback by substituting mutual information with the notion of directed information in some cases as in continuous time developed in [6].

The rest of the paper is organized as follows. Section II describes the concrete problem setting. In Section III, we present and discuss the main results. We refer the reader to [7] for proofs of results and more thorough discussions. In Sections IV and V, we provide examples and simulation results.

II. Problem Setting

Let the input process $X^T = \{X_t, 0 \leq t \leq T\}$ be governed by probability law P_θ from some class of possible laws indexed by $\theta \in \mathcal{S}$. $\mathcal{S}$ is an uncertainty set known to the estimator. Let Y^T be the noise corrupted observations of X^T at the estimator, therefore, the probability law of Y^T also depends on the particular realization of $\theta \in \mathcal{S}$. Denote the input and reconstruction alphabets by $\mathcal{X}$ and $\hat{\mathcal{X}}$, respectively. In other words, $X_t \in \mathcal{X}$ and $\hat{X}_t \in \hat{\mathcal{X}}$, where typically both $\mathcal{X}$ and $\hat{\mathcal{X}}$ are $\mathbb{R}$ or $\mathbb{R}_+$ for each t. Let the measurable[1] $l(\cdot,\cdot) : \mathcal{X} \times \hat{\mathcal{X}} \mapsto [0,\infty)$ be a given loss function. For simplicity and transparency of our arguments, we assume that $\hat{\mathcal{X}}$ is a convex set and that $l(\cdot,\cdot)$ satisfies the following properties:

(P1) $l(x,\hat{x})$ is convex in $\hat{x}$ for each x;
(P2) $\inf_{\hat{x}\in\hat{\mathcal{X}}} \mathbb{E}[l(X,\hat{x})] = \mathbb{E}[l(X,\mathbb{E}[X])]$

The squared error loss function and the natural loss function $l(x,\hat{x}) = x\log(\frac{x}{\hat{x}}) - x + \hat{x}$, introduced in [3], are examples of loss functions satisfying this property. Cf. [8] for other loss functions of this type.

Define the causal estimator $\hat{X}_t(\cdot)$ as a function of the output process up to time t, i.e. $Y^t = \{Y_s, 0 \leq s \leq t\}$ and also define the causal estimation error associated with the filter $\hat{X} = \{\hat{X}_t(\cdot), 0 \leq t \leq T\}$ by

$$\text{cmle}(\theta, \hat{X}) = \mathbb{E}_{P_\theta}\left[\int_0^T l(X_t, \hat{X}_t(Y^t))dt\right]$$

where $E_{P_\theta}[\cdot]$ denotes expectation under P_θ.

III. Main Results

A. Minimax Causal Estimation Criterion

Suppose the estimator is optimized for law Q while the active law is P_θ. Then the estimator will employ the Bayesian estimator $\hat{X}_Q$, where $\hat{X}_Q = \{\hat{X}_{Q,t}(\cdot) : \hat{X}_{Q,t}(Y^t) = \mathbb{E}_Q[X_t|Y^t], 0 \leq t \leq T\}$ denotes the Bayesian filter under prior Q, and the corresponding mismatched causal estimation error will be

$$\text{cmle}(\theta, \hat{X}_Q) = \mathbb{E}_{P_\theta}\left[\int_0^T l(X_t, \mathbb{E}_Q[X|Y^t])dt\right] \triangleq \text{cmle}_{\theta,Q}.$$

In particular, when the estimator is optimized for the true distribution, i.e., $Q = P_\theta$, the causal estimation error is

$$\text{cmle}(\theta, \hat{X}_{P_\theta}) = \mathbb{E}_{P_\theta}\left[\int_0^T l(X_t, \mathbb{E}_{P_\theta}[X|Y^t])dt\right] = \text{cmle}_{\theta,P_\theta}.$$

Because of (P2), this is the Bayes optimum for the source P_θ.

Clearly, this can be considered our benchmark because it is the minimum causal estimation error when the probability law is exactly known. Now, similar to the universal source coding problem, define the regret of the filter $\hat{X}$ when the active source is P_θ by

$$R(\theta, \hat{X}) = \text{cmle}(\theta, \hat{X}) - \text{cmle}_{\theta,P_\theta}.$$

Since $\text{cmle}_{\theta,P_\theta}$ is our benchmark, it is natural to seek to minimize the worst-case regret over all possible $\theta \in \mathcal{S}$. Specifically, define minimax($\mathcal{S}$) as

$$\text{minimax}(\mathcal{S}) = \inf_{\hat{X}} \sup_{\theta\in\mathcal{S}} R(\theta, \hat{X}),$$

where the infimum is over all possible filters.

B. Main Results

Similar to (1), if the estimator is Bayesian under law Q, i.e., $\hat{X}_t(Y^t) = \mathbb{E}_Q[X_t|Y^t]$, then denote the regret by

$$R(\theta, \hat{X}) \triangleq R_{\theta,Q}.$$

Theorem 1: Let $\mathcal{Q}$ denote the convex hull of the uncertainty set of all probability laws, i.e. $\mathcal{Q} = \text{conv}(\{P_\theta; \theta \in \mathcal{S}\})$. Let $l(\cdot,\cdot)$ be a loss function with the above properties. Then

$$\begin{aligned}\text{minimax}(\mathcal{S}) &= \min_{Q\in\mathcal{Q}} \sup_{\theta\in\mathcal{S}} R_{\theta,Q} \\ &= \min_{Q\in\mathcal{Q}} \sup_{\theta\in\mathcal{S}} \{\text{cmle}_{\theta,Q} - \text{cmle}_{\theta,P_\theta}\}. \qquad (1)\end{aligned}$$

Consider the following two canonical continuous-time channel models.

1) Gaussian Channel: Suppose that under all P_θ, $\theta \in \mathcal{S}$, Y^T is the AWGN corrupted version of X^T, i.e.,

$$dY_t = X_t dt + dW_t$$

where W^T is standard Brownian motion independent of X^T. We consider half the squared loss function which is $l(x,\hat{x}) = \frac{1}{2}(x-\hat{x})^2$, where we introduce the factor $1/2$ to streamline the exposition that follows.

[1] From this point on we tacitly assume measurability of all functions introduced.

2) Poisson Channel: Suppose that under all P_θ, $\theta \in \mathcal{S}$, Y^T is a non-homogeneous Poisson process with intensity X^T, where X^T is a stochastic process bounded by two positive constants. As in [3], we employ the natrual loss function $l(x, \hat{x}) = x \log(x/\hat{x}) - x + \hat{x}$. This loss function is a natural choice for the Poisson channel, cf. [3, Lemma 2.1].

Note that in these two settings the uncertainty in P_θ is only in the distribution of X^T, as the channel from X^T to Y^T is the same regardless of θ. We are now ready to state our main results.

Theorem 2 (Regret-Capacity): Let the setting be either that of the Gaussian channel or the Poisson channel. Then

$$\text{minimax}(\mathcal{S}) = \sup_{w \in \mu(\mathcal{S})} I(\Theta; Y^T) \tag{2}$$

where $\mu(\mathcal{S})$ denotes the class of all possible measures on the set $\mathcal{S}$. $I(\Theta; Y^T)$ denotes the mutual information between Θ and Y^T where $\Theta \sim w$ is a random variable on $\mathcal{S}$ and the conditional law of Y^T given $\Theta = \theta$ is the law of Y^T under P_θ.

Theorem 3 (Minimax Filter): Suppose the supremum in Theorem 2 is achieved and let w^* denote the achiever. Then the minimum in (1) is achieved by the Bayesian optimal filter with respect to Q^*, the mixture of P_θ's with respect to w^*, i.e.,

$$Q^* = \int_{\theta \in \mathcal{S}} P_\theta w^*(d\theta)$$

and the minimax filter is

$$\hat{X}_t(Y^t) = \mathbb{E}_{Q^*}[X_t | Y^t].$$

Theorem 4 (Strong Regret-Capacity): Suppose the supremum in Theorem 2 is achieved and let w^* denote the achiever. For any filter $\hat{X}$ and every $\epsilon > 0$,

$$R(\theta, \hat{X}) > (1 - \epsilon) \cdot \text{minimax}(\mathcal{S})$$

for all $\theta \in \mathcal{S}$ with the possible exception of points in a subset $B \subset \mathcal{S}$, where

$$w^*(B) \leq e \cdot 2^{-\epsilon \cdot \text{minimax}(\mathcal{S})}.$$

Consider the case of the presence of feedback. Suppose X_t is also affected by previous output $\{Y_s : 0 \leq s < t\}$. Let $\mathcal{P}$ be a class of joint laws of X^T, Y^T and $\mathcal{S}$ be a set of indices of laws. Definition of minimax and $R_{\theta,Q}$ remain the same. Then, above theorems also hold, i.e.,

Theorem 5 (Presence of Feedback):

$$\text{minimax}(\mathcal{S}) = \min_{Q \in \mathcal{Q}} \sup_{\theta \in \mathcal{S}} R_{\theta,Q}$$

Moreover, if the setting is either Gaussian or Poisson, then

$$\begin{aligned} \text{minimax}(\mathcal{S}) &= \min_{Q \in \mathcal{Q}} \sup_{\theta \in \mathcal{S}} R_{\theta,Q} \\ &= \sup_w I(\Theta; Y^T) \\ &= \sup_w I(X^T \to Y^T) - I(X^T \to Y^T | \Theta) \end{aligned}$$

where $I(X^T \to Y^T)$ is the directed information from X^T to Y^T, as introduced in [6].

C. Discussion

Theorem 1 implies that the optimum minimax filter is a Bayesian filter under some law Q. Furthermore, this minimum achieving Q is a mixture of P_θ's. Therefore, in order to find the optimum minimax filter, it is enough to restrict the search space to that of Bayesian filters. This is equivalent to finding an optimal prior Q^*, or optimum weights w^* over laws $\{P_\theta\}$. Note that we have not assumed anything on the statistics of the input and output processes but only the above mentioned properties of the loss function $l(\cdot, \cdot)$.

Theorem 2 implies that there is a strong link between the minimax regret and the communication problem, as in the theory of universal source coding. This mutual information is equal to $I(X^T; Y^T) - I(X^T; Y^T | \Theta)$ where the first term is the mutual information between input and output when the input distirbution is $Q = \int_\theta P_\theta w(d\theta)$. Furthermore, Theorem 3 provides a prescription for such a filter in cases where the noise corruption mechanism is either Gaussian or Poisson. Note that if the uncertainty set consists of a set that constrains the possible underlying signals rather than their laws (e.g., all signals X^T at the channel input confined to some peak and or power constraint) then the right hand side of (2) boils down to a supremum over all distributions on the set of allowable channel inputs, i.e.,

$$\begin{aligned} \text{minimax}(\mathcal{S}) &= \sup_{w \in \mu(\mathcal{S})} I(X^T; Y^T) \\ &= \sup_{P_{X^T} \in \mathcal{Q}} I(X^T; Y^T), \end{aligned} \tag{3}$$

where $\mathcal{Q} = \text{conv}(\mathcal{P})$. (3) follows because X^T is deterministic given Θ, therefore, $I(X^T; Y^T | \Theta) = 0$.

Note that the right hand side of the above equation is the capacity of the channel whose input is constrained to lie in the uncertainty set of signals at the channel input with respect to which the minimax quantity is defined. Moreover, letting Q^* denote the capacity achieving distribution, the optimum minimax estimator is the Bayesian estimator with respect to the law Q^*. More interestingly, Q^* turns out to coincide with the classical notion of the least favorable prior from estimation theory. We would refer to [7, Appendix I] for this connection in detail. These results show the strong relation between the minimax estimation and channel coding problems.

In Theorem 4, we can see that our optimal minimax estimator minimizes not only the worst case regret, but also the regret for most $\theta \in \mathcal{S}$ under distribution w^*. Cf. [4] for a discussion of the significance and implications of this result. For example, it implies that when $\mathcal{S}$ is a compact subset of $\mathbb{R}^k$ and the parametrization of the input distributions P_θ is sufficiently smooth, the minimax filter is essentially optimal not only in the worst case sense for which it was optimized, but in fact on "most" of the sources over all possible filters (Note that we are not restricting filters to be Bayesian). "Most" here means that the Lebesgue measure of the set of parameters indexing sources for which is vanishing as the value of $\text{minimax}(\mathcal{S})$ is growing without bound, which is usually the case as T increases in all but the most degenerate of situations.

Theorem 5 implies that the above result can be extended to the case where feedback exists. Note that if $\mathcal{P}$ is a class of deterministic laws, i.e, X_t is a function of previous inputs and outputs, then,

$$\text{minimax}(\mathcal{S}) = \sup_{w} I(X^T \to Y^T).$$

IV. Examples

A. Gaussian Channel and Sparse Signal

Based on the above theorems, we first apply them to the problem of sparse signal estimation under Gaussian noise.

1) Setting: We assume output process Y^T is AWGN corrupted version of X^T as we discussed in Section III-B1, while input process X^T is sparse which will be explained in the following. Recall that we are using half of a mean squared error as a distortion measure, $l(x, \hat{x}) = \frac{1}{2}(x - \hat{x})^2$.

Let $\{\phi_i(t), 0 \le t \le T\}_{i=1}^n$ be a given orthonormal signal set. Suppose X^T is a linear combination of $\phi_i(t)$'s, i.e. $X_t = \sum_{i=1}^n A_i \phi_i(t)$ where $\{A_i\}_{i=1}^n$ are random variables with unknown distribution. However, we assume that the estimator knows that the signal X^T is power constrained and is sparse, by which we mean that the fraction of non-zero elements in $\{A_i\}$ should be smaller than q (i.e., at most nq number of A_i's can be nonzero). Let $\mathcal{P}$ be a class of all possible probability measures P_θ of vector $A = (A_1, \cdots, A_n)$ indexed by θ which satisfies these two constraints, i.e.,

$$\mathcal{P} = \left\{ P_\theta : \frac{1}{n}\sum_{i=1}^n A_i^2 \le P, \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{A_i \neq 0\}} \le q \text{ a.s.} \right\}. \quad (4)$$

Note that $\int_0^T X_t^2 dt = \sum_{i=1}^n A_i^2$ because of orthonormality of basis, therefore, it is equivalent to consider $\frac{1}{n}\sum_{i=1}^n A_i^2 \le P$ as a power constraint. Define an uncertainty set $\mathcal{S}$ by set of such indices. It is clear that $\mathcal{P} = \{P_\theta : \theta \in \mathcal{S}\}$ is a convex set.

We further define $\mathcal{P}_D$ as a class of deterministic measures $P_\theta \in \mathcal{P}$ (i.e., $P_\theta(\{a^n\}) = 1$ for some $a^n \in \mathbb{R}^n$), and the corresponding set of indices as $\mathcal{S}_D$. Note that $\text{conv}(\mathcal{P}_D) = \mathcal{P}$.

Our goal is to find a minimax estimator and $\text{minimax}(\mathcal{S})$ where $\mathcal{S}$, which represents the set of power constrained sparse input signals, is given.

2) Apply the Theorem: Theorem 2 implies that

$$\text{minimax}(\mathcal{S}) = \sup_{w(\cdot) \in \mu(\mathcal{S})} I(X^T; Y^T) - I(X^T; Y^T|\Theta).$$

Since our optimum causal minimax estimator is Bayesian estimator under the distribution $Q^* = \int P_\theta w^*(d\theta)$ where w^* is supremum achiever, we are interested in w^*. Rather than maximizing the difference between mutual informations, we can find an equivalent problem which is much easier to handle by exploiting the relation between $\text{minimax}(\mathcal{S})$ and $\text{minimax}(\mathcal{S}_D)$. Moreover, the minimum achiever Q^* of $\text{minimax}(\mathcal{S}_D)$ coincides with that of $\text{minimax}(\mathcal{S})$.

Lemma 6:

$$\text{minimax}(\mathcal{S}_D) = \text{minimax}(\mathcal{S}).$$

Since $\mathcal{P}_D$ is a set of deterministic measures, we can get more explicit formula of $\text{minimax}(\mathcal{S}_D)$ as we showed in Section III-C,

$$\text{minimax}(\mathcal{S}) = \text{minimax}(\mathcal{S}_D) = \sup_{P_\theta \in \mathcal{P}} I(X^T; Y^T).$$

Since X^T is governed by the law $\int P_\theta w(d\theta)$, therefore, it is equivalent to maximize the mutual information over all possible mixture law instead of finding optimum measure on $\mathcal{S}_D$.

3) Sufficient Statistics: Since the channel input signal is a linear combination of orthonormal signals, sufficient statistics of the channel output signal are projections on each ϕ_i's, i.e., $\{\int_0^T \phi_i(t) dY_t\}_{i=1}^n$. Therefore, the above mutual information $I(X^T; Y^T)$ can be further simplified as

$$\text{minimax}(\mathcal{S}) = \sup_{P_\theta \in \mathcal{P}} I\left(A^n; B^n\right).$$

where $B_i = \int_0^T \phi_i(t) dY_t$ for $1 \le i \le n$. Since we assumed an orthonormal basis, B^n can be viewed as the output of a discrete-time additive white Gaussian channel, i.e. $B_i = A_i + W_i$ where W_i is i.i.d. standard Gaussian noise and independent of A^n. This implies that our problem of maximizing the mutual information over the continuous time channel is equivalent to maximizing the mutual information between n channel inputs and n channel outputs over the AWGN channel, with the input distribution constrained as in (4).

Note that $\left\{\int_0^t \phi_i(s) dY_s\right\}_{i=1}^n$ is still sufficient statistics of outputs for all $t < T$. This allows us to construct the causal filter $\mathbb{E}[X_t|Y^t]$ much more easily.

4) Gaussian Channel with Sparsity Constraint: Let define the class of sparse signals with average constraints

$$\mathcal{P}_{av} = \left\{ P_\theta : \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n A_i^2\right] \le P, \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{A_i \neq 0\}}\right] \le q \right\}.$$

and the corresponding index set $\mathcal{S}_{av}$. The problem $\sup_{P_\theta \in \mathcal{P}_{av}} I(A^n; B^n)$ was recently considered by Zhang and Guo in [9], where they referred to it as "Gaussian channels with duty cycle and power constraints". They have shown that the distribution on A^n that maximizes this mutual information is i.i.d. and discrete. In other words, letting P_d denote the distribution on A that maximizes $I(A; B)$, when $B = A + W$ for a standard Gaussian noise W which is independent of A, among all distributions constrained by $\mathbb{E}[A^2] \le P$ and $P(A \neq 0) \le q$, their results imply that P_d is discrete. Thus, we can argue that

$$\sup_{P_\theta \in \mathcal{P}_{av}} I(A^n; B^n) = n\left[I(A; B)\right]_{P_A = P_d}.$$

5) Bayesian Estimator: Let Q^* be the minimum achieving law of $\text{minimax}(\mathcal{S})$ so that the optimum causal minimax estimator is a Bayesian estimator assuming the prior Q^*, i.e.,

$$\hat{X}_t(Y^t) = \mathbb{E}_{Q^*}[X_t|Y^t] = \mathbb{E}_{Q^*}\left[X_t \,\middle|\, \left\{\int_0^t \phi_i(s) dY_s\right\}_{i=1}^n\right].$$

6) Almost Optimal Causal Minimax Estimator: It is hard to find a maximum achieving distribution in some cases, indeed most of the problems of finding capacity achieving distribution are still open including our sparse signal estimation problem. Therefore, we will use an approximated version of the prior, $\tilde{Q}$, so that we can easily implement the filter. One natural choice of $\tilde{Q}$ is the capacity achieving distribution of $\sup_{P_\theta \in \mathcal{P}_{av}} I(A;B)$ which is i.i.d. of P_d. Indeed, the corresponding worst case regret is close to the optimum, which is minimax$(\mathcal{S})$. We refer to [7] for rigorous proof.

B. Poisson Channel and Direct Current Signal

Consider direct current (DC) signal estimation over the Poisson channel. The input process $X_t \equiv X$ for all $0 \le t \le T$, where X is a random variable bounded by $a \le X \le A$ where a, A are positive constants. We can define uncertainty set $\mathcal{S}$ such that $\{P_\theta : \theta \in \mathcal{S}\}$ is the set of all possible probability measures on X under which $a \le X \le A$ almost surely. The estimator observes Poisson process with rate X_t and performance is measured under the natural log loss function $l(x, \hat{x}) = x \log(x/\hat{x}) - x + \hat{x}$.

Since $\{P_\theta : \theta \in \mathcal{S}\}$ is convex and since Y_T is a sufficient statistic of Y^T for X^T (which is constant at X), we have

$$\text{minimax}(\mathcal{S}) = \sup I(X; Y_T),$$

where the maximization is over all distributions on X supported on $[a, A]$. Corresponding communication problem is that of the capacity of the discrete-time poisson channel, where the input is non-negative, real valued X with a peak power constraint $a \le X \le A$ a.s. and the output is Poisson random variable with parameter TX. In this scenario, Shamai [10] showed that capacity achieving distribution is discrete with finite number of mass points. Let P_s be this capacity achieving distribution. Although analytic expression of P_s and capacity of the channel are still open, we can approximate the distribution numerically to arbitrary precision.

Using Theorem 3, we can conclude that the optimum minimax causal estimator is conditional expectation of X given Y_t with respect to the distribution P_s, i.e.,

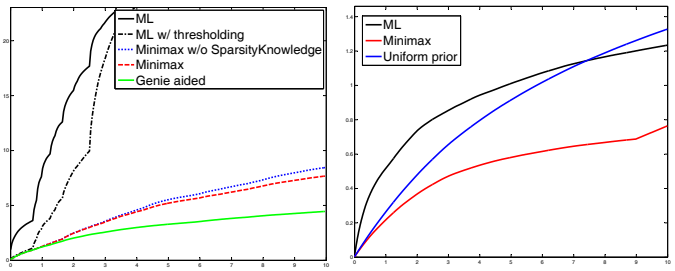
$$\hat{X}_t(Y^t) = \mathbb{E}_{P_s}[X|Y_t].$$

V. Experiments

A. Gaussian Channel and Sparse Signal

Consider the setting of Section IV-A. In order to compare the performance of the suggested minimax filter, we introduce some possible estimators, which are maximum likelihood (ML) estimator, the minimax estimator that lacks the sparsity information and genie aided scheme which allows additional information of source. We assume that genie aided scheme knows the exact postions of nonzeros of source

Similar to [9], we approximate P_d with finite number of mass points. Using approximated version of P_d, we compare the performance of estimator in Figure 1a. Here we set $n = 7$, $k = 2$, $P = 10^{0.4}$ (4dB), and Haar basis as an orthonormal signal set. We generate random sparse coefficient and take an average of causal squared error over 100 simulations. Note that we are randomly generated signals therefore causal errors in the above experiments are not the worst case error, however, we can check that optimum minimax estimator outperforms maximum likelihood estimators and minimax estimator without sparsity knowledge.



(a) Gaussian Channel (b) Poisson Channel

Fig. 1: Plots of cmle for the experiment of SectionV-A and V-B. Here we have taken $T = 10$ for both cases. X_t is randomly generated over 100 times and we computed average causal loss for each filter.

B. Poisson Channel and DC Signal

For comparison, we present some other natural estimators, which are ML estimator and Bayesian estimator with uniform prior.

Figure 1b shows numerical results for $a = 0.5$, $A = 2$ case. We take an average of causal mean loss error over 100 times for $X = 0.5, 1, 1.5, 2$ and plot an worst case error.

References

[1] T. Duncan, "On the Calculation of Mutual Information," *SIAM Journal on Applied Mathematics*, vol. 19, no. 1, pp. 215–220, 1970.

[2] T. Weissman, "The Relationship Between Causal and Noncausal Mismatched Estimation in Continuous-Time AWGN Channels," *Information Theory, IEEE Transactions on*, vol. 56, no. 9, pp. 4256–4273, Sep. 2010.

[3] R. Atar, T. Weissman, "Mutual Information, Relative Entropy, and Estimation in the Poisson Channel," *Information Theory, IEEE Transactions on*, vol. 58, no. 3, pp. 1302 –1318, Mar. 2012.

[4] N. Merhav, M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *Information Theory, IEEE Transactions on*, vol. 41, no. 3, pp. 714–722, May 1995.

[5] J. Rissanen, "Universal coding, information, prediction, and estimation," *Information Theory, IEEE Transactions on*, vol. 30, no. 4, pp. 629–636, July 1984.

[6] T. Weissman, Y.-H. Kim, and H. Permuter, "Directed information, causal estimation, and communication in continuous time," *Information Theory, IEEE Transactions on*, vol. PP, no. 99, p. 1, Nov. 2012.

[7] A. No, T. Weissman, "Minimax Filtering Regret via Relations Between Information and Estimation," *available at http://arxiv.org/abs/1301.5096*.

[8] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a Bregman predictor," *Information Theory, IEEE Transactions on*, vol. 51, no. 7, pp. 2664–2669, July 2005.

[9] L. Zhang, D. Guo, "Capacity of Gaussian channels with duty cycle and power constraints," *in Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on,* Aug. 2011, pp. 513 –517.

[10] S. Shamai, "On the capacity of a direct-detection photon channel with intertransition-constrained binary input," *Information Theory, IEEE Transactions on*, vol. 37, no. 6, pp. 1540 –1550, Nov. 1991.