On the Submodularity of Linear Experimental Design

Matthias Seeger
Probabilistic Machine Learning and Medical Image Processing
Saarland University
Room 116, Campus E1.4, 66123 Saarbruecken
mseeger@mmci.uni-saarland.de

May 19, 2009

Abstract

Here, I review facts that are most probably known, namely that the information gain criterion used to drive experimental design in a linear-Gaussian model is submodular, so that a well-known approximation guarantee holds for the sequential greedy algorithm. The criterion is equal to a certain mutual information, which is not submodular in general. I point out the high potential relevance of obtaining approximation guarantees for nonlinear experimental design as well.

1 Submodularity of Linear Experimental Design

Let $\boldsymbol{u} \in \mathbb{R}^n$ a latent vector of interest, $\boldsymbol{X} \in \mathbb{R}^{M\times n}$ a (complete) design matrix, $\boldsymbol{r} = \boldsymbol{X}\boldsymbol{u}$, and $\boldsymbol{y} = \boldsymbol{r} + \boldsymbol{\varepsilon}$, where $\boldsymbol{u} \sim P(\boldsymbol{u}) = N(\boldsymbol{0}, \boldsymbol{I})$ and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ independently. Given a subset $I \subset \{1, \dots, M\}$, we are interested in reconstructing $\boldsymbol{u}$ from measurements $\boldsymbol{y}_I$ obtained with the design $\boldsymbol{X}_{I,\cdot} \in \mathbb{R}^{|I|\times n}$.

The goal of experimental design is to choose a subset I, so that the posterior uncertainty in $\boldsymbol{u}|\boldsymbol{y}_I$ is as small as possible, over all subsets of the same size. The criterion of interest is

$$f(I) := \mathrm{H}[P(\boldsymbol{u})] - \mathrm{E}_{\boldsymbol{y}_I}[\mathrm{H}[P(\boldsymbol{u}|\boldsymbol{y}_I)]] = \mathrm{H}[\boldsymbol{u}] - \mathrm{H}[\boldsymbol{u}|\boldsymbol{y}_I] = \mathrm{I}(\boldsymbol{u}; \boldsymbol{y}_I).$$

It is well known that $f(I)$ is nondecreasing (since conditioning reduces entropy), even if $P(\boldsymbol{u})$ is not Gaussian. Also, $f(\emptyset) = 0$. Moreover, for $j \notin I$,

$$f(I \cup \{j\}) - f(I) = \mathrm{H}[\boldsymbol{u}|\boldsymbol{y}_I] - \mathrm{H}[\boldsymbol{u}|\boldsymbol{y}_I, y_j] = \mathrm{H}[r_j|\boldsymbol{y}_I] - \mathrm{H}[r_j|\boldsymbol{y}_I, y_j],$$

because clearly $P(\boldsymbol{u} \setminus r_j | r_j, \boldsymbol{y}_I) = P(\boldsymbol{u} \setminus r_j | r_j, \boldsymbol{y}_I, y_j)$. Note that all variables here are jointly Gaussian. Here and below, I use two properties of Gaussians. First, the entropy of a Gaussian depends on the covariance matrix only, and second, the covariance matrix of a conditional Gaussian distribution $P(\boldsymbol{a}|\boldsymbol{b})$ does not depend on the value of $\boldsymbol{b}$. Therefore, if $\rho_j(I) := \mathrm{Var}[r_j|\boldsymbol{y}_I]$, this variance does not depend on $\boldsymbol{y}_I$ (but of course on $\boldsymbol{X}_{I,\cdot}$). Moreover, $P(r_j|\boldsymbol{y}_I, y_j) \propto P(r_j|\boldsymbol{y}_I)N(y_j|r_j, \sigma^2)$, so that

$$2f(I \cup \{j\}) - 2f(I) = \log \rho_j(I) + \log[\sigma^{-2} + \rho_j(I)^{-1}] = \log[1 + \sigma^{-2}\rho_j(I)].$$

All that remains to show is that for any $I_1 \subset I_2$ and $j \notin I_2$, $\rho_j(I_1) \geq \rho_j(I_2)$ (because $\log(1+\cdot)$ is increasing). But $\rho_j(I) = \psi(\mathrm{H}[r_j|\boldsymbol{y}_I])$ with ψ strictly increasing, so the result follows by "conditioning reduces entropy".

This argument depends strongly on the peculiar property of Gaussians, because only if $\rho_j(I)$ is independent of $\boldsymbol{y}_I$, and monotonically related to $\mathrm{H}[r_j|\boldsymbol{y}_I]$, can we express $f(I \cup \{j\}) - f(I)$ as a monotonic function of $\rho_j(I)$ only.

2 Extensions to Nonlinear Experimental Design?

Unfortunately, experimental design for a linear-Gaussian model (also called linear experimental design) is not very interesting in practice, except possibly the Gaussian prior $P(\boldsymbol{u})$ is fitted to independent data gathered otherwise (which is somewhat unnatural: shouldn't this prior data sampling be optimized as well?). Note that Gaussian process ED scenarios with Gaussian likelihood are just a special case of a linear-Gaussian model, as long as parameters such as the covariance function or hyperparameters are fitted beforehand, not depending on the data sampled during the design optimization.

For example, a Gaussian prior is a poor description of natural or medical images, and applications such as the sampling optimization of magnetic resonance imaging are driven much better by a non-Gaussian (in this case a sparse) linear model [1]. Beyond images, most real-world signals are poorly described by Gaussians, partly explaining why ICA works for a wide range of signals. For temporal data (such as audio streams), the Gaussian assumption is highly unnatural, because significant jumps (which occur in about any real temporal data *somewhere*) are drastically penalized. In general, if signal representations are modelled using sparsity potentials rather than Gaussians, a boost of performance is observed often in practice. In my opinion, theory supporting linear experimental design is of rather limited impact, just because linear experimental design is of limited usefulness in practice, while any theoretical understanding of *nonlinear* ED properties would have direct consequences in practice for a number of high-profile applications, among them (adaptive) compressed sensing, acquisition optimization, or computational photography.

What would be required to get a handle into the nonlinear ED case? The main difficulty is that in this case, the optimal design depends on what is in fact measured (the $\boldsymbol{y}_I$). While the formulation above is closed-loop, in that $\mathrm{H}[\boldsymbol{u}|\boldsymbol{y}_I]$ does not depend on $\boldsymbol{y}_I$, which is integrated out w.r.t. the model, this is not realistic in scenarios such as [1], where real measurements (from "nature") are obtained at every design extensions. While it is also interesting to analyze the closed-loop scenario with non-Gaussian $P(\boldsymbol{u})$, this means that design decisions are taken entirely based on a model which is known to reflect the truth poorly (as a generative model), and also that there is nothing adaptive about such a procedure (in that no real-world training data is used, except maybe to fit the prior beforehand). While this is the "active learning" world then, which seems different from the usual combinatorial-vs-greedy setting in which submodularity seems to be used, it would be highly important to connect them in a reasonable way.

A start could be made by focussing on priors $P(\boldsymbol{u})$ that can be written as scale mixtures of Gaussians, hoping that some of the Gaussian properties can still be used. In fact, this is not a large restriction at all. The prior used in [1] is

$$P(\boldsymbol{u}) = \mathrm{E}_{\boldsymbol{\gamma}}[N(\boldsymbol{u}|\boldsymbol{0}, \sigma^2[\boldsymbol{B}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{B}]^{-1})$$

with some distribution $P(\boldsymbol{\gamma}) = \prod_i P(\gamma_i)$, which does not depend on $\boldsymbol{u}$. If $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$, this prior factorizes w.r.t. the s_i. Posteriors are scale mixtures as well,

$$P(\boldsymbol{u}|\boldsymbol{y}_I) = \mathrm{E}_{\boldsymbol{\gamma}}[N(\boldsymbol{u}|\boldsymbol{h}, \sigma^2\boldsymbol{A}^{-1})], \quad \boldsymbol{A} = \boldsymbol{X}_{I,\cdot}^T\boldsymbol{X}_{I,\cdot} + \boldsymbol{B}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{B},\ \boldsymbol{h} = \boldsymbol{A}^{-1}\boldsymbol{X}_{I,\cdot}^T\boldsymbol{y}_I,$$

which however do not factorize anymore. Now, things become messy. For example,

$$\mathrm{H}[\boldsymbol{u}|\boldsymbol{y}_I] = \mathrm{E}\left[-\log \mathrm{E}_{\boldsymbol{\gamma}}[N(\boldsymbol{u}|\boldsymbol{h}, \sigma^2\boldsymbol{A}^{-1})]\right],$$

which cannot be computed analytically. Also, there is no direct relationship to $\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}_I]$, and the latter is also not a simple expression:

$$\mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}_I] = \sigma^2\mathrm{E}_{\boldsymbol{\gamma}}[\boldsymbol{A}^{-1}] + \mathrm{Cov}_{\boldsymbol{\gamma}}[\boldsymbol{A}^{-1}\boldsymbol{X}_{I,\cdot}^T\boldsymbol{y}_I].$$

On the other hand, a number of variational bounds on such expressions are known.

The final, and maybe toughest, problem for lifting theory to nonlinear ED in this context is that posterior inference is not done exactly, but using (say) variational approximations. Instead of $P(\boldsymbol{u}|\boldsymbol{y}_I)$, a Gaussian approximation $Q(\boldsymbol{u}|\boldsymbol{y}_I)$ is used. In the relaxation used in [1], the scale parameters $\boldsymbol{\gamma}$ above becomes *variational parameters* (with different semantics). With previous methods such as expectation propagation, it is not even clear how to characterise the final $Q(\boldsymbol{u}|\boldsymbol{y}_I) \approx P(\boldsymbol{u}|\boldsymbol{y}_I)$, because it might not be unique (given $\boldsymbol{y}_I$). The relaxation in [1] is proven to be convex: for some convex criterion $\phi(\boldsymbol{\gamma}|\boldsymbol{y}_I)$, *the* approximation $Q(\boldsymbol{u}|\boldsymbol{y}_I) = Q(\boldsymbol{u}|\boldsymbol{y}_I;\boldsymbol{\gamma}_*)$ is given by $\boldsymbol{\gamma}_* = \mathrm{argmin}_{\boldsymbol{\gamma}}\,\phi$.

How would theory look like in this nested case, where the evaluation of $f(I)$ entails some inner (convex) optimization? In my opinion, this is the real theoretical challenge, simply because this is what people do in practice. Posteriors cannot be computed exactly, and variational relaxations are increasingly used. In this context, a strong point is made about convexity of an inference relaxation, because that should simplify and robustify higher-level decision problems. But how does theory look like to really make this point? In the framework of [1], at least $\phi(\boldsymbol{\gamma}|\boldsymbol{y}_I)$ is known explicitly, and is of a rather simple form (and convex). But what do such properties mean for the maximization of $f(I)$?

References

[1] M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf. Bayesian experimental design of magnetic resonance imaging sequences. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1441–1448. MIT Press, 2009.