Sample-then-optimize posterior sampling for Bayesian linear models

Alexander G. de. G Matthews
University of Cambridge [*]

Jiri Hron
University of Cambridge

Richard E. Turner
University of Cambridge

Zoubin Ghahramani
University of Cambridge, Uber AI Labs

1 Introduction

In modern machine learning it is common to train models which have an extremely high intrinsic capacity. The results obtained are often initialization dependent, are different for disparate optimizers and in some cases have no explicit regularization. This raises difficult questions about generalization [1]. A natural approach to questions of generalization is a Bayesian one. There is therefore a growing literature attempting to understand how Bayesian posterior inference could emerge from the complexity of modern practice [2, 3], even without having such a procedure as the stated goal.

In this work we consider a simple special case where exact Bayesian posterior sampling emerges from sampling (c.f initialization) and then gradient descent. Specifically, for a Bayesian linear model, if we parameterize it in terms of a deterministic function of an isotropic normal prior, then the action of sampling from the prior followed by first order optimization of the squared loss will give a posterior sample. Although the assumptions are stronger than many real problems, it still exhibits the challenging properties of redundant model capacity and a lack of explicit regularizers, along with initialization and optimizer dependence. It is therefore an interesting controlled test case. Given its simplicity, the method itself may turn out to be of independent interest from our original goal. Whilst the material of Section 2 is classical [4], the material in Section 3 is, as far as we are aware, novel.

2 Problem specification

Consider a design matrix Φ with M columns and N rows. For our purposes M can be thought of as the number of features and N the number of data points. Let w be a column vector of size M which will contain feature weights. Let y be column vector of size N which will represent targets. In this work we will be interested in analyzing the following model

$$w \sim \mathcal{N}(0, \lambda I) \tag{1}$$

$$y = \Phi w \tag{2}$$

[*] Contact by email at *am554@cam.ac.uk*

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Such models can be classified into three cases. In the first case, there is no solution for w and the inverse problem is ill-specified. In the second case, there is a unique solution for w and the posterior distribution is a unit point mass on this solution. In the third case, there is a whole space of solutions and the posterior is non-trivial. We will assume case three in all that follows. Although the problem specification may seem limiting, in fact a broad variety of normal linear models can be written in this form. Sampling from any normal distribution can be reduced to standard normal sampling followed by an affine transformation. This affine transformation can be represented by suitable definitions of y and Φ. Further, we can represent noisy observations by augmenting the weight vector w with additional weights that only affect single components of y. The general solution to the problem is given by:

$$\hat{w} = \Phi^P y + (I - \Phi^P \Phi)\epsilon. \tag{3}$$

Here Φ^P is the Moore-Penrose pseudo inverse (or just pseudo inverse in what follows) of the design matrix. I is the identity and ϵ is a general column vector of size M, which allows us to cover the space of solutions. The term $(I - \Phi^P \Phi)$ is interpretable as a projection onto the null space of Φ. The null space must be not be empty for there to be multiple solutions. The linear constraint in (1) can be associated with an optimization objective:

$$f(w) = \|\Phi w - y\|_2^2 \,. \tag{4}$$

By assumption, there will exist many weight vectors such that $f(w)$ is zero. We could form a regularized version of the objective function.

$$g(w) = f(w) + \sigma^2 w^T w. \tag{5}$$

The solution is then unique and has a closed form given by

$$\tilde{w} = (\Phi^T \Phi + \sigma^2 I)^{-1} \Phi^T y. \tag{6}$$

There are two limit relations for the pseudo inverse, which exist even if the unregularized inverses do not:

$$\Phi^P = \lim_{\sigma^2 \to 0^+} \left\{ \Phi^T \Phi + \sigma^2 I \right\}^{-1} \Phi^T = \lim_{\sigma^2 \to 0^+} \Phi^T \left\{ \Phi \Phi^T + \sigma^2 I \right\}^{-1}. \tag{7}$$

The first limit relation relates the regularized solution (6) to the unregularized one (3).

3 Sample then optimize posterior sampling

Our proposed method for obtaining exact posterior samples from the model in Equation (1) is to draw a sample w_0 from the prior then optimize the weights to convergence by first order gradient descent under the *unregularized* objective in Equation (4). This gives exact posterior samples from the model as we now detail.

To start with, we analyze the gradient descent dynamics. The change in the weights Δw for first order gradient descent is given by:

$$\Delta w = -k \left(\Phi^T \Phi w - \Phi^T y \right), \tag{8}$$

where k represents a constant step size parameter. We consider the action of this update by viewing the vector space $\mathbb{R}^M$ as the Cartesian product of the null space of Φ and its orthogonal complement. The update does not change the component of the weight vector w that lies in the null space. To see this note that $e^T \Phi^T f = f^T \Phi e = 0$ for any pair of test vectors $e \in \mathbb{R}^M$, $f \in \mathbb{R}^N$, with e in the null space. Recall that $(I - \Phi^P \Phi)$ represents projection into the null space. The component of w_0 in the null space is thus $(I - \Phi^P \Phi) w_0$, and this remains unchanged during optimization.

Next, we turn our attention to the behavior of the component $w_{||}$ of the weight vector w which lies in the orthogonal complement of the null space. There is a fixed point given by:

$$\Phi^T \Phi w_{||} = \Phi^T y. \tag{9}$$

In this subspace, the solution for $w_{||}$ is now unique and is given by $\Phi^P y$. From an arbitrary initialization the fixed point will be attained if k is sufficiently small that the dynamics do not oscillate indefinitely or exponentially diverge. Under this assumption, the limiting solution w^* attained by the optimization dynamics will be

$$w^* = \Phi^P y + (I - \Phi^P \Phi) w_0 \tag{10}$$

The proposed sampling of w_0 from the prior distribution induces a normal distribution on the limiting solution:

$$w^* \sim \mathcal{N}(\Phi^P y, \lambda(I - \Phi^P \Phi)) \tag{11}$$

Having analyzed the distribution induced on the attained solution of the optimization problem, we now show that this matches the Bayesian posterior distribution for the model in Equation (1). The only slightly non-standard aspect of the necessary solution is that the noiseless observations mean that the joint normal distribution of w and y can be degenerate. We thus perform the analysis assuming normal noise on y with variance σ^2 and then take the limit $\sigma^2 \to 0^+$. The conditional distribution in question can be found in a number of standard works (see for instance the Appendix of Rasmussen and Williams [5]):

$$w^*|y \sim \mathcal{N}(\Phi^T(\Phi\Phi^T + \sigma^2 I)^{-1} y, \lambda I - \lambda \Phi^T (\Phi\Phi^T + \sigma^2 I)^{-1} \Phi) \tag{12}$$

Taking the limit $\sigma^2 \to 0^+$ and using the relation for the pseudo inverse from Equation (7) we obtain the claimed equivalence with the distribution (11). It is natural to ask if we can generalize the method to sampling from any prior covariance matrix. We have found using counter examples that this is not the case, though see Section 2 for our discussion of reparameterization.

4 Application to Gaussian process regression

In this section, we demonstrate the inference method using a simple one dimensional Gaussian process (GP) regression task. For clarity of exposition we will assume no observation noise, although as previously stated, this assumption could be relaxed. The model is:

$$f \sim \mathcal{GP}(0, K) \tag{13}$$

with $\mathcal{GP}$ denoting a Gaussian process and K denoting the covariance function which we take to be an radial basis function kernel with length scale $1/2$ and unit signal variance. We use a simple basis function approximation to give a distribution over functions $\tilde{f}$:

$$w \sim \mathcal{N}(0, I) \tag{14}$$

$$\tilde{f}(x) = \sum_{i=1}^{M} w_i \phi_i(x) \tag{15}$$

where ϕ_i are the basis functions. Here we take 50 evenly spaced Gaussian basis functions in the range $[-10, 10]$. Whilst more sophisticated GP approximations exist, this will be sufficient to demonstrate the sample-then-optimize inference scheme. To make the connection to the model (1) take N input/output pairs $(x_j, y_j)_{j=1}^N$ and the relation $y(x_j) = \tilde{f}(x_j)$. For more detail see Rasmussen and Williams, Chapter 2 [5]. We also compared to exact Gaussian process regression with the model (13). The results are shown in Figure 1. The basis function approximation error is negligible in this example. The alignment between the two methods is very close as predicted by our theory. It is also interesting to note how each prior function sample moves to become a posterior function sample (of matching color) in the top two sub-figures. With such a flexible prior, it is of course possible to achieve zero loss in a number of ways. The movement in the functions that we actually obtain is a property of the optimizer dynamics. The sample functions change more closer to the data points. Indeed, far from the data they hardly move from their prior values. This provides one intuition for how generalization could emerge in flexible models trained greedily.

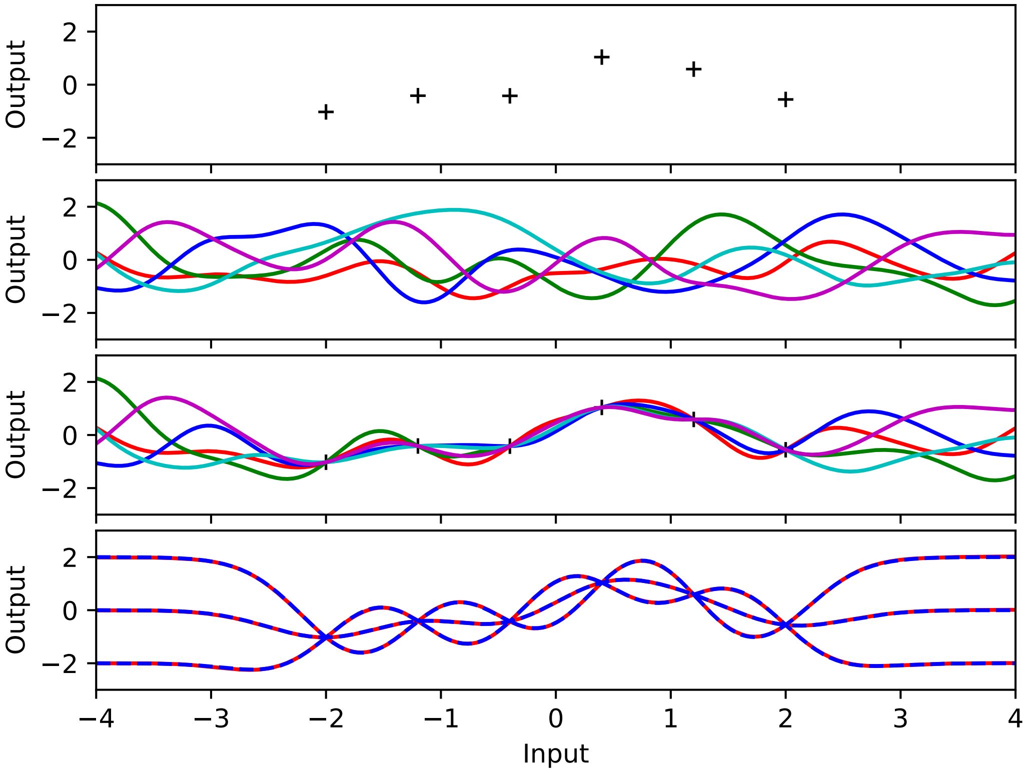

Figure 1: An illustration of the method for Gaussian process regression. Top: A simple regression data set. Middle Upper: Draws from the basis function prior defined by Equation (14). Middle Lower: The result of first order gradient descent run to convergence on each of the models. The colors match those of the prior samples. These functions will be exact posterior samples from the approximate model (14). Bottom: The mean and 2σ-credible interval of ten thousand such samples (red) compared to exact Gaussian process regression with the exact prior (13) (blue dashed). The two methods match very closely.

5 Conclusions

In this work we have shown that for a broad class of Bayesian linear models posterior sampling can emerge from prior sampling followed by deterministic first order gradient descent. We argued in the introduction that such a special case, despite its tractability, is nevertheless an interesting test case for understanding how Bayesian induction could arise in more complex models.

With this in mind the question arises as to how far such an approach can be generalized. It would have been useful if any form of normal sample could be used in the sampling phase because this would allow exact online inference. The question of less straightforward generalizations will occupy us in further work. In any case, the Gaussian process example, along with the extreme simplicity of the approach, suggest that even without further additions the method may be of practical use.

References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding Deep Learning requires rethinking Generalization. *International Conference on Learning Representations (ICLR)*, 2016.

[2] Max Welling and Yee Whye Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. *International Conference on Machine Learning (ICML)*, 2011.

[3] Stephan Mandt, Matthew Hoffman, and David Blei. A Variational Analysis of Stochastic Gradient Algorithms. *International Conference on Machine Learning (ICML)*, 2016.

[4] E. H. Moore. On the Reciprocal of the General Algebraic Matrix. *Bulletin of the American Mathematical Society*, 1920.

[5] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.