Early Stopping as Nonparametric Variational Inference

David Duvenaud [†]
Harvard University

Dougal Maclaurin [†]
Harvard University

Ryan P. Adams
Harvard University

Abstract

We show that unconverged stochastic gradient descent can be interpreted as a procedure that samples from a nonparametric approximate posterior distribution. This distribution is implicitly defined by the transformation of an initial distribution by a sequence of optimization steps. By tracking the change in entropy over these distributions during optimization, we form a scalable, unbiased estimate of a variational lower bound on the log marginal likelihood. This bound can be used to optimize hyperparameters instead of cross-validation. This Bayesian interpretation of SGD suggests improved, overfitting-resistant optimization procedures, and gives a theoretical foundation for early stopping and ensembling. We investigate the properties of this marginal likelihood estimator on neural network models.

1 Introduction

We propose an interpretation of incomplete optimization in terms of variational Bayesian inference, and provide a simple method for estimating the marginal likelihood of the approximate posterior. Our starting point is a Bayesian posterior distribution for a potentially complicated model, in which there is an empirical loss that can be interpreted as a negative log likelihood and regularizers that have interpretations as priors. One might proceed with MAP inference, and perform an optimization to find the best parameters. The main idea of this paper is that such an optimization procedure, initialized according to some distribution that can be chosen freely, generates a sequence of distributions that are implicitly defined by the action of the optimization update rule on the previous distribution. We can treat these distributions as approximations to the true posterior distribution. A single optimization run for N iterations represents a draw from the Nth such distribution in the sequence. Figure 1 shows contours of these approximate distributions on an example posterior.

With this interpretation, the number of optimization iterations can be seen as a variational parameter, one that trades off fitting the data well against maintaining a broad (high entropy) distribution. Early stopping amounts to optimizing the variational lower bound (or an approximation based on a validation set) with respect to this variational parameter. Ensembling different random restarts can be viewed as taking independent samples from the variational posterior.

1.1 Contributions

- We introduce a new interpretation of optimization algorithms as samplers from a variational distribution that adapts to the true posterior, eventually collapsing around its modes.
- We provide a scalable estimator for the entropy of these implicit distributions, allowing us to estimate a lower bound on the marginal likelihood of any model whose posterior is twice-differentiable, even on problems with millions of parameters and data points.
- We investigate the performance of these estimators empirically on neural network models, and show that they have reasonable properties.

[†] Equal contributors.

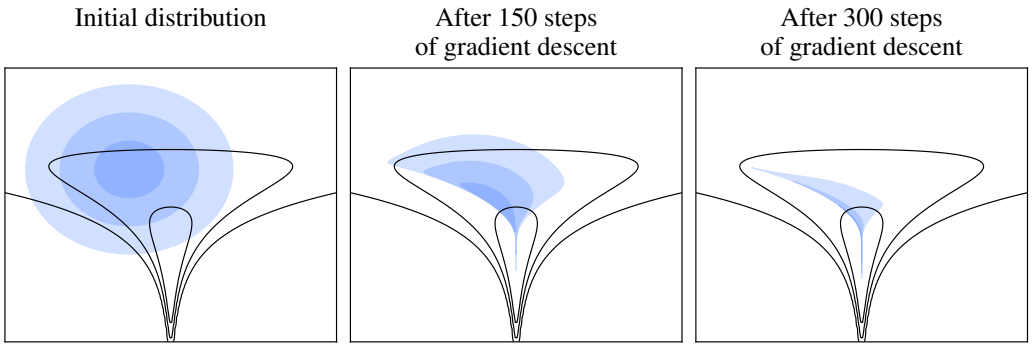

Figure 1: A series of distributions (blue) implicitly defined by gradient descent on an objective (black). These distributions are defined by mapping each point in the initial distribution through a fixed number of iterations of optimization. These distributions have nonparametric shapes, and eventually concentrate around the optima of the objective.

2 Incomplete optimization as variational inference

Variational inference aims to approximate an intractable posterior distribution, $p(\theta|\mathbf{x})$, with another more tractable distribution, $q(\theta)$. The usual measure of the quality of the approximation is the Kullback-Leibler (KL) divergence from $q(\theta)$ to $p(\theta, \mathbf{x})$. This measure provides a lower bound on the marginal likelihood of the original model; applying Bayes' rule to the definition of $\mathrm{KL}\left(q(\theta)\|p(\theta|\mathbf{x})\right)$ gives the familiar inequality:

$$\log p(\mathbf{x}) \geq \underbrace{-\mathbb{E}_{q(\theta)}\left[-\log p(\theta, \mathbf{x})\right]}_{\text{Energy } E[q]} \underbrace{-\mathbb{E}_{q(\theta)}\left[\log q(\theta)\right]}_{\text{Entropy } S[q]} := \mathcal{L}[q] \tag{1}$$

To perform variational inference, we require a family of distributions over which to maximize $\mathcal{L}[q]$. Consider a general procedure to minimize the energy $(-\log p(\theta, \mathbf{x}))$ with respect to $\theta \in \mathbb{R}^D$. The parameters θ are initialized according to some distribution $q_0(\theta)$ and updated at each iteration according to a transition operation $T : \mathbb{R}^D \to \mathbb{R}^D$:

$$\theta_0 \sim q_0(\theta) \tag{2}$$

$$\theta_{t+1} = T(\theta_t) \tag{3}$$

Our variational family consists of the sequence of distributions $q_0, q_1, q_2, \ldots$, where $q_t(\theta)$ is the distribution over θ_t generated by the above procedure. These distributions don't have a closed form, but we can exactly sample from q_t by simply running the optimizer for t steps starting from a random initialization.

We cannot evaluate $\mathcal{L}[q_t]$ exactly, but we can obtain an unbiased estimator. Sampling θ_0 from q_0 and then applying the transition operator t times produces an exact sample θ_0 from q_t, by definition. Since θ_t is an exact sample from $q_t(\theta)$, $\log p(\theta_t, \mathbf{x})$ is an unbiased estimator of the energy term of (1). The entropy term is trickier, since we do not have access to the density $q(\theta)$ directly. However, if we know the entropy of the initial distribution, $S[q_0(\theta)]$, then we can estimate $S[q_t(\theta)]$ by tracking the change in entropy at each iteration, calculated by the change of variables formula.

To compute how the volume shrinks or expands due to an iteration of the optimizer, we require access to the Jacobian of the optimizer's transition operator, $J(\theta)$:

$$S[q_{t+1}] - S[q_t] = \mathbb{E}_{q_t(\theta_t)}\left[\log |J(\theta_t)|\right]. \tag{4}$$

Note that this analysis assumes that the mapping T is bijective. Combining these terms, we have an unbiased estimator of $\mathcal{L}$ at iteration T, based on the sequence of parameters, $\theta_0, \ldots, \theta_T$, from a single training run:

$$\mathcal{L}[q_T] \approx \underbrace{\log p(\theta_T, \mathbf{x})}_{\text{Energy}} + \underbrace{\sum_{t=0}^{T-1} \log |J(\theta_t)| + S[q_0]}_{\text{Entropy}}. \tag{5}$$

The entropy of stochastic gradient descent Stochastic gradient descent is a popular optimization procedure with the following update rule:

$$\theta_{t+1} = \theta_t - \alpha \nabla L(\theta), \tag{6}$$

where the $L(\theta)$ the objective loss (or an unbiased estimator of it e.g. using minibatches) for example $-\log p(\theta, \mathbf{x})$, and α is a 'step size' hyperparameter. Taking the Jacobian of this update rule gives the following unbiased estimator for the change in entropy at each iteration:

$$S[q_{t+1}] - S[q_t] \approx \log |I - \alpha H_t(\theta_t)| \tag{7}$$

where H_t is the Hessian of $-\log p_t(\theta, \mathbf{x})$ with respect to θ.

Estimating the Jacobian in high dimensions The expression for the change in entropy given by (7) is impractical for large-scale problems since it requires an $\mathcal{O}\left(D^3\right)$ determinant computation. Fortunately, we can make a good approximation using just two Hessian-vector products, which can usually be performed in $\mathcal{O}\left(D\right)$ time using reverse-mode differentiation (Pearlmutter, 1994). The idea is that since $\alpha\lambda_{\text{max}}$ is small, the Jacobian is just a small perturbation to the identity, and we can approximate its determinant using traces as follows:

$$\log |I - \alpha H| = \sum_{i=0}^{D} \log\left(1 - \alpha\lambda_i\right) \quad \geq \sum_{i=0}^{D} \left[-\alpha\lambda_i - (\alpha\lambda_i)^2\right] = -\alpha \text{Tr}\left[H\right] - \alpha^2 \text{Tr}\left[HH\right] . \tag{8}$$

The bound in (8) is just a second order Taylor expansion of $\log(1 - x)$ about $x = 0$ and is valid if $\alpha\lambda_i < 0.68$. However, gradient descent is unstable anyway if $\alpha\lambda_{\text{max}} > 2$, where λ_{max} is the largest eigenvalue of H_t. So choosing a conservative learning rate keeps this bound in the correct direction. For sufficiently small learning rates, this bound becomes tight.

The trace of the Hessian can be estimated using inner products of random vectors (Bai et al., 1996):

$$\text{Tr}\left[H\right] = \mathbb{E}\left[\mathbf{r}^T H \mathbf{r}\right], \qquad \mathbf{r} \sim \mathcal{N}(0, I) . \tag{9}$$

Initialization objective functions What initial parameter distribution should we use for SGD? The marginal likelihood estimate given by (5) is valid no matter which initial distribution we choose. We use the prior as the initial distribution and log-likelihood as the objective in our experiments.

3 Experiments

Here, we demonstrate that our marginal likelihood estimator has reasonable properties on two tasks. Code for all experiments is available at `github.com/HIPS/maxwells-daemon`.

Choosing the number of training iterations We performed regression on the Boston housing dataset using a neural network with one hidden layer having 100 hidden units, sigmoidal activation functions, and no regularization. Figure 2 shows that marginal likelihood peaks at a similar place to the peak of held-out log-likelihood, which is where early stopping would occur when using a large validation set.

Choosing the number of hidden units Figure 3 shows marginal likelihood estimates as a function of the number of hidden units in the hidden layer of a neural network trained on 50,000 MNIST handwritten digits. The largest network trained in this experiment contains 2 million parameters.

4 Limitations

Using only a single sample to estimate both the expected likelihood as well as the entropy of an entire distribution will necessarily have high variance under some circumstances. These problems could conceivably be addressed by ensembling, which has an interpretation as taking multiple exact independent samples from the implicit posterior.

Second, as parameters converge, their entropy estimate (and true entropy) will continue to decrease indefinitely, making the marginal likelihood arbitrarily small. However, in practice there is usually

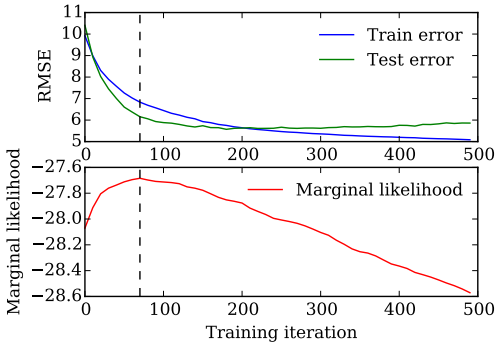

Figure 2: *Top*: Training and test-set error on the Boston housing dataset. *Bottom*: Stochastic gradient descent marginal likelihood estimates. The dashed line indicates the iteration with highest marginal likelihood. The marginal likelihood, estimated online using only the training set, and the test error peak at a similar number of iterations.

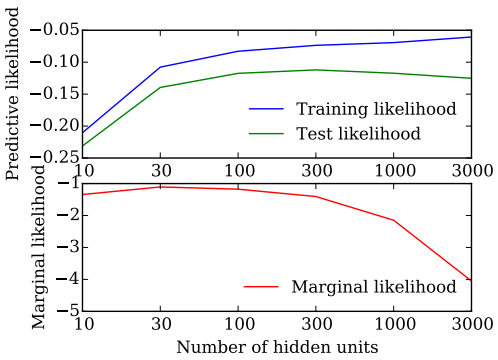

Figure 3: *Top*: Training and test-set likelihood as a function of the number of hidden units in the first layer of a neural network. *Bottom*: Stochastic gradient descent marginal likelihood estimates. In this case, the marginal likelihood over-penalizes high numbers of hidden units.

a limit to the degree of overfitting possible, making our marginal likelihood estimate a poor guide to generalization performance.

Finally, figure 1 shows that the distribution implied by SGD collapses to a small portion of the true posterior early on, and continues to shrink as optimization proceeds. However, the point of early stopping is not that the intermediate distributions are particularly good approximations, but simply that they are better than the point masses that occur when optimization has converged.

5 Related work

Our method can be seen as a special case of Salimans et al. (2014), who showed that any set of stochastic dynamics, even those not satisfying detailed balance, can be used to implicitly define a variational distribution. Hardt et al. (2015) give theoretical results showing that the smaller the number of training epochs, the better the generalization performance of models trained using SGD. Toulis et al. (2015) examine the properties of SGD as an estimator, and show that a variant that averages parameter updates has improved statistical efficiency.

One possible method to deal with over-zealous reduction in entropy by SGD would be to add noise to the dynamics. In the case of Gaussian noise, we would recover Langevin dynamics (Neal, 2011).

References

Bai, Zhaojun, Fahey, Gark, and Golub, Gene. Some large-scale matrix computation problems. *Journal of Computational and Applied Mathematics*, 74(1):71–89, 1996.

Hardt, Moritz, Recht, Benjamin, and Singer, Yoram. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

Neal, Radford M. MCMC using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2, 2011.

Pearlmutter, Barak A. Fast exact multiplication by the Hessian. *Neural computation*, 6(1):147–160, 1994.

Salimans, Tim, Kingma, Diederik P., and Welling, Max. Markov chain Monte Carlo and variational inference: Bridging the gap. *arXiv preprint arXiv:1410.6460*, 2014.

Toulis, Panos, Tran, Dustin, and Airoldi, Edoardo M. Stability and optimality in stochastic gradient descent. *arXiv preprint arXiv:1505.02417*, 2015.