Early Stopping is Nonparametric Variational Inference

David Duvenaud*, Dougal Maclaurin*, Ryan Adams

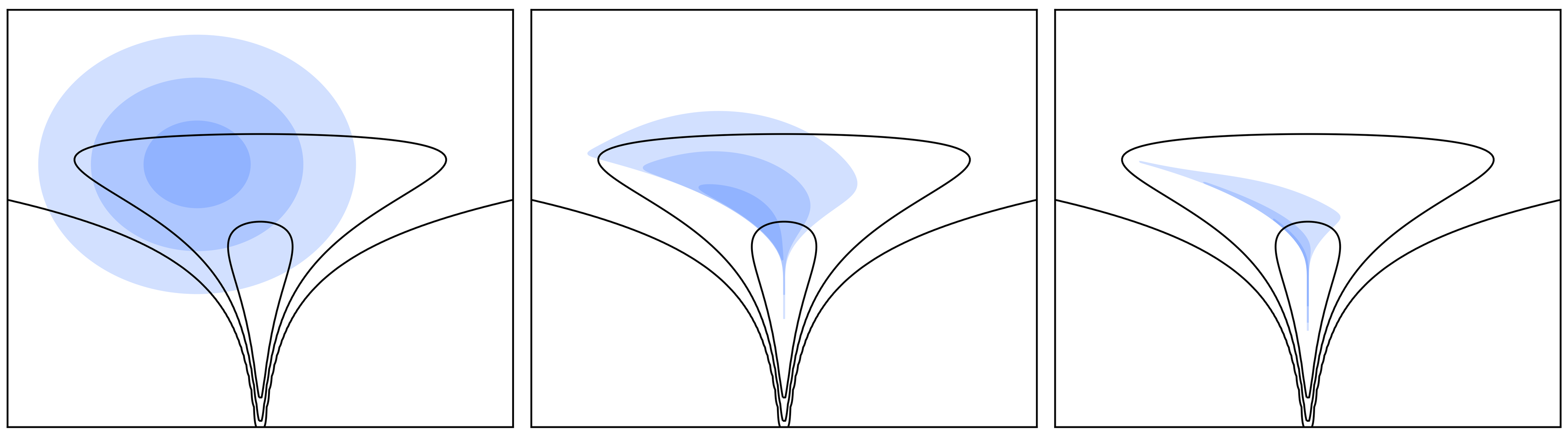
Why does early stopping help?

Regularization = MAP inference
Limiting model capacity = Bayesian Occam's razor
Cross-validation = Estimating marginal likelihood
Dropout = Integrating out spike-and-slab
Ensembling = Bayes model averaging?
Early stopping = ??

Gradient descent with random starts is a sampler

What is the implicit distribution of parameters after optimizing for t steps?

Initial distribution | After 150 steps | After 300 steps



Distributions (blue) implicitly defined by gradient descent on an objective (black).

- Starts as a bad approximation (prior dist)
- Ends as a bad approximation (point mass)
- Ensembling = taking multiple samples from dist
- Early stopping = choosing best intermediate dist

Cross validation vs. marginal likelihood

- What if we could evaluate marginal likelihood of implicit distribution?
- Could choose all hypers to maximize marginal likelihood
- No need for cross-validation?

Contribution: Variational Lower Bound

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

Energy estimated from optimized objective function (training loss is NLL):

$$\mathbb{E}_{q(\theta)}\left[-\log p(\theta,\mathbf{x})\right] \approx -\log p(\hat{\theta}_T,\mathbf{x})$$

Entropy estimated by tracking change at each iteration:

$$-\mathbb{E}_{q(\theta)}\left[\log q(\theta)\right] \approx S[q_0] + \sum_{t=0}^{T-1} \log\left|J(\hat{\theta}_t)\right|$$

Using a single sample!

Estimating change in entropy

- Inuitively: High curvature makes entropy decrease quickly
- Can measure local curvature with Hessian
- Approximation good for small step-sizes

Volume change given by Jacobian of optimizer's operator:

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

Gradient descent update rule:

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

Has Jacobian:

$$J(\theta_t) = I - \alpha\nabla\nabla L(\theta_t)$$

Entropy change estimated at a single sample:

$$S[q_{t+1}] - S[q_t] \approx \log\left|I - \alpha\nabla\nabla L(\theta_t)\right|$$

SGD with entropy estimate

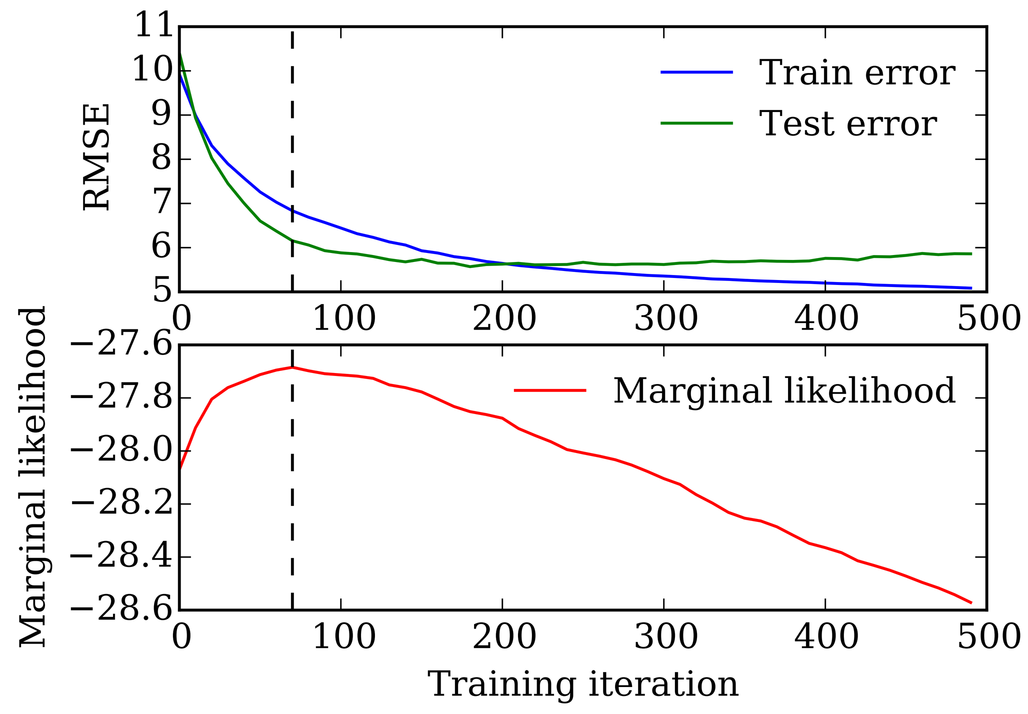
input: Weight init scale σ_0, step size α, negative log-likelihood $L(\theta,t)$
initialize $\theta_0 \sim \mathcal{N}(0,\sigma_0\mathbf{I}_D)$
initialize $S_0 = \frac{D}{2}(1+\log 2\pi) + D\log\sigma_0$
for $t=1$ **to** T **do**
  $S_t = S_{t-1} + \log\left|\mathbf{I} - \alpha\nabla\nabla L(\theta_t,t)\right|$
  $\theta_t = \theta_{t-1} - \alpha\nabla L(\theta_t,t)$
end for
output sample θ_T, entropy estimate S_T

Computational Complexity

- Approximate bound: $\log p(\mathbf{x}) \gtrsim -L(\theta_T) + S_T$
- Determinant is $\mathcal{O}(D^3)$
- $\mathcal{O}(D)$ Taylor approximation using Hessian-vector products
- Scales linearly in parameters and dataset size

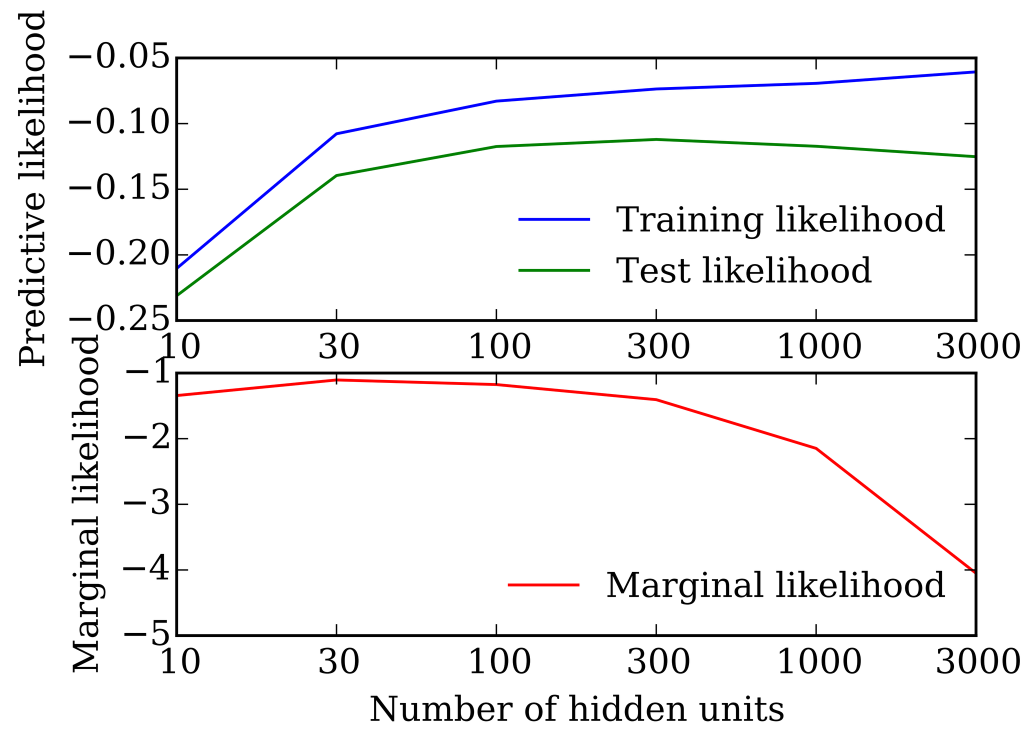
Example: Choosing when to stop

- Neural network on the Boston housing dataset.
- SGD marginal likelihood estimate gives stopping criterion without a validation set



Example: Choosing number of hidden units

- Neural net on 50000 MNIST digits
- Largest model has 2 million params
- Gives reasonable estimates, but cross-validation still better
- Entropy bound over-penalizes after long training



Main Takeaways

- Optimization with random restarts implies nonparametric intermediate dists
- Early stopping chooses among these distributions
- Ensembling samples from them
- Can scalably estimate lower bound on model evidence during optimization
- Bound can be used for Langevin-dynamics recognition networks!
- All code at `github.com/HIPS/maxwells-daemon`