CS 156a Learning Systems - Homework #3

Ling Li, ling@cs.caltech.edu

November 7, 2000

3.1 Bias and Variance

For learning model $\{g_\alpha(x)\}$, the mean-squared-error of hypothesis $g_\alpha(x)$ is

$$E(\alpha) = \mathcal{E}_x\left[\left(g_\alpha(x) - f(x)\right)^2\right] = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(g(x) - f(x)\right)^2 dx.$$

Thus part (a) of the problem is to find the combination of parameters that minimizes $E(\alpha)$; Part (b) is to compute $\hat{g}(x) = \mathcal{E}_T\left[g_T(x)\right]$ with respect to the training set T; And in part (c), the out-of-sample error on any point $x \in [-\pi, \pi]$ is

$$\mathcal{E}_T\left[\left(g_T(x) - f(x)\right)^2\right] = \sigma^2\left[g_T(x)\right] + \left(\hat{g}(x) - f(x)\right)^2,$$

with bias $\left(\hat{g}(x) - f(x)\right)^2$ and variance

$$\sigma^2\left[g_T(x)\right] = \mathcal{E}_T\left[\left(g_T(x) - \hat{g}(x)\right)^2\right] = \mathcal{E}_T\left[g_T^2(x) - \hat{g}^2(x)\right].$$

(i) $g(x) = ax + b$, $f(x) = \sin x$.

(a) $E = \dfrac{1}{2\pi}\displaystyle\int_{-\pi}^{\pi}\left(ax + b - \sin x\right)^2 dx.$

$$\begin{aligned}
\frac{\partial E}{\partial a} &= \frac{1}{\pi}\int_{-\pi}^{\pi}\left(ax + b - \sin x\right) x dx = \frac{2}{3}\left(a\pi^2 - 3\right), \\
\frac{\partial E}{\partial b} &= \frac{1}{\pi}\int_{-\pi}^{\pi}\left(ax + b - \sin x\right) dx = 2b.
\end{aligned}$$

Since the Hessian of E

$$H = \begin{pmatrix} \frac{\partial^2 E}{\partial a^2} & \frac{\partial^2 E}{\partial a \partial b} \\ \frac{\partial^2 E}{\partial b \partial a} & \frac{\partial^2 E}{\partial b^2} \end{pmatrix} = \begin{pmatrix} \frac{2\pi^2}{3} & 0 \\ 0 & 2 \end{pmatrix}$$

is a positive-definite matrix, E gets to its minimum at $a = \frac{3}{\pi^2} \approx 0.303964$ and $b = 0$, where $\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0$, and the best hypothesis is $g(x) = \frac{3}{\pi^2}x \approx 0.303964x$.

(b) For any set of two examples $(x', \sin x')$ and $(x'', \sin x'')$, the hypothesis that minimizes the mean squared error on these samples is the line that goes through them. That is,[1]

$$g_T(x) = \frac{\sin x' - \sin x''}{x' - x''}(x - x'') + \sin x''.$$

Thus

$$\hat{g}(x) = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left[\frac{\sin x' - \sin x''}{x' - x''}(x - x'') + \sin x''\right]dx'dx'' \approx 0.246986x.$$

(c) The bias is

$$(\hat{g}(x) - f(x))^2 \approx (0.246986x - \sin x)^2 \tag{1}$$

and the variance is

$$\begin{aligned}
\sigma^2\left[g_T(x)\right] &= \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left[\frac{\sin x' - \sin x''}{x' - x''}(x - x'') + \sin x''\right]^2 dx'dx'' - \hat{g}^2(x) \\
&= \left[\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left(\frac{\sin x' - \sin x''}{x' - x''}\right)^2 dx'dx'' - \left(\frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\frac{\sin x' - \sin x''}{x' - x''}dx'dx''\right)^2\right]x^2 \\
&\quad + \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\left(\frac{x'\sin x'' - x''\sin x'}{x' - x''}\right)^2 dx'dx'' \\
&\approx 0.232634x^2 + 0.910947. \qquad (2)
\end{aligned}$$

Thus the out-sample-error is

$$(\hat{g}(x) - f(x))^2 + \sigma^2\left[g_T(x)\right] \approx (0.246986x - \sin x)^2 + 0.232634x^2 + 0.910947.$$

Figure 1(a) shows these values.

(ii) $g(x) = b$, $f(x) = \sin x$.

(a) $E = \frac{1}{2\pi}\int_{-\pi}^{\pi}(b - \sin x)^2\,dx = \frac{1}{2} + b^2$. E gets to its minimum at $b = 0$. So the best hypothesis is $g(x) = 0$.

(b) For any set of two examples $(x', \sin x')$ and $(x'', \sin x'')$, the hypothesis that minimizes the mean squared error on them is

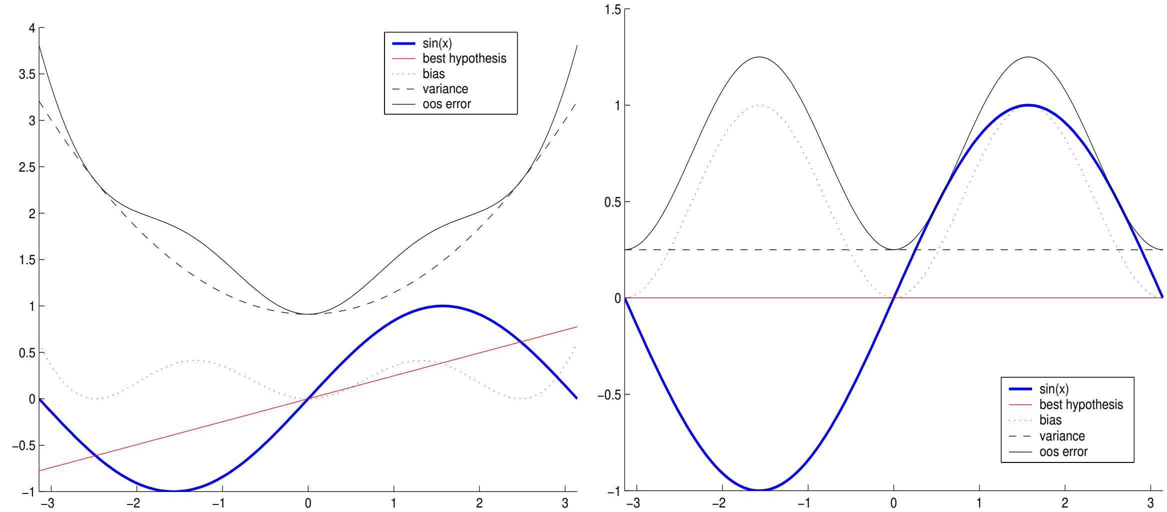
$$g(x) = \frac{\sin x' + \sin x''}{2}.$$

Thus

$$\hat{g}(x) = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\frac{\sin x' + \sin x''}{2}dx'dx'' = 0.$$

[1] For the infinitesimal-probability case of two identical examples, we assume

$$\frac{\sin x' - \sin x''}{x' - x''} \triangleq \frac{d\sin x'}{x'} = \cos x'.$$



(a) $g(x) = ax + b$ (b) $g(x) = b$

Figure 1: Bias and variance components for $f(x) = \sin x$, $x \in [-\pi, \pi]$, and learning models $\{ax + b\}$ and $\{b\}$.

(c) The bias is

$$(\hat{g}(x) - f(x))^2 = \sin^2 x$$

which is bigger than (1). The variance is

$$\sigma^2 [g_T(x)] = \frac{1}{4\pi^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \left(\frac{\sin x' + \sin x''}{2} \right)^2 dx' dx'' - \hat{g}^2(x) = \frac{1}{4}$$

which is smaller than (2). The out-sample-error is

$$(\hat{g}(x) - f(x))^2 + \sigma^2 [g_T(x)] = \sin^2(x) + \frac{1}{4}.$$

Figure 1(b) shows these values.

3.2 Gradient and Hessian

(i) From homework #1,

$$\begin{aligned} \frac{\partial E}{\partial u} &= 2 \left(u e^{v} - 2 v e^{-u} \right) \left(e^{v} + 2 v e^{-u} \right), \\ \frac{\partial E}{\partial v} &= 2 \left(u e^{v} - 2 v e^{-u} \right) \left(u e^{v} - 2 e^{-u} \right). \end{aligned}$$

Thus the Hessian at $(u, v) = (1, 1)$ is

$$H = \begin{pmatrix} 2e^2 + 4 + 16e^{-2} & 4e^2 - 16e^{-2} \\ 4e^2 - 16e^{-2} & 4e^2 - 12 + 8e^{-2} \end{pmatrix} \approx \begin{pmatrix} 20.9435 & 27.3909 \\ 27.3909 & 18.6389 \end{pmatrix}.$$

(ii) If we descend along the gradient, i.e.,

$$\nabla E = \left(\frac{\partial E}{\partial u}, \frac{\partial E}{\partial v} \right)^T,$$

$$\begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} = -0.1 \frac{\nabla E}{|\nabla E|},$$

the second-order approximation is

$$\Delta_2 E = (\nabla E)^T \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} + \frac{1}{2} (\Delta u, \Delta v) \, H \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} \approx -1.358985.$$

The first-order calculated in homework #1 is

$$\Delta_1 E \approx -1.579104$$

and the exact value is

$$\Delta E \approx -1.372005.$$

Thus $\Delta_1 E < \Delta E < \Delta_2 E$ and, compared to $\Delta_1 E$, $\Delta_2 E$ is closer to the exact value ΔE.

(iii) Computing the values of $E(u, v)$ at 20,000 evenly distributed points in a circle centered at $(1, 1)$ with radius 0.1, we get numerically the minimum of ΔE is

$$\Delta E \approx -1.376419$$

when $\Delta u = -0.09074844$ and $\Delta v = -0.04200857$.

(iv) Using second-order approximation and repeating (iii), we get the minimum of $\Delta_2 E$ is

$$\Delta_2 E \approx -1.366360$$

when $\Delta u = -0.09192530$ and $\Delta v = -0.03959943$.

(v) Neglecting the off-diagonal elements of the Hessian H and repeating (iv), we get the minimum of $\Delta_2' E$ is

$$\Delta_2' E \approx -1.477273$$

when $\Delta u = -0.08640816$ and $\Delta v = -0.05033518$.

3.3 Momentum and Adaptive Learning Rate

$$E(\mathbf{w}) = \left(\sum_i e^{a_i w_i} + \sum_i e^{-b_i w_i} \right) \times \left(\sum_i e^{c_i w_i} + \sum_i e^{-d_i w_i} \right).$$

$$\begin{aligned} \frac{\partial E}{\partial w_i} &= \left(a_i e^{w_i} - b_i e^{-b_i w_i} \right) \times \left(\sum_i e^{c_i w_i} + \sum_i e^{-d_i w_i} \right) \\ &\quad + \left(\sum_i e^{a_i w_i} + \sum_i e^{-b_i w_i} \right) \times \left(c_i e^{c_i w_i} - d_i e^{-d_i w_i} \right). \end{aligned}$$

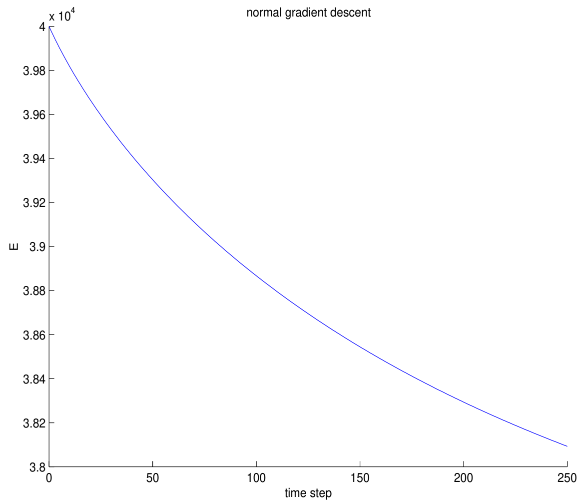


Figure 2: Minimize $E(\mathbf{w})$ with gradient descent. Start with $\mathbf{w} = 0$ and learning rate $\eta = 10^{-5}$.

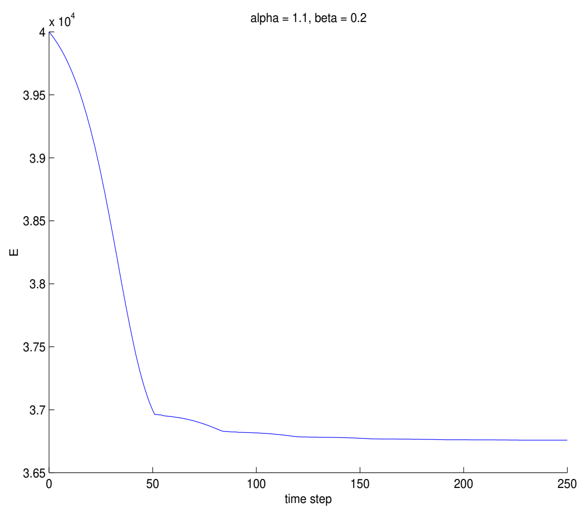


Figure 3: Minimize E with adaptive learning rate: $\eta = 10^{-5}$, $\alpha = 1.1$ and $\beta = 0.2$.

(i) Starting with $\mathbf{w}_0 = 0$, the normal gradient descent (Figure 2) gives E at time step 250 is 3.80929×10^4.

(ii) The best combination of parameters seems to be $\alpha = 1.1$ and $\beta = 0.2$. For this combination (Figure 3), E at time step 250 is 3.67581×10^4, which is smaller than the E in (i), and the error curve of E v.s. t falls more rapidly in Figure 3 than in Figure 2.

(iii) $\mu = 0.99$ gives the best performance (Figure 4): E at time step 250 is 3.69317×10^4. Compared to (i), this E is also smaller, and the curve also falls more rapidly. However, the curve fluctuates during learning and the E at time step 250 is bigger than that in (ii).

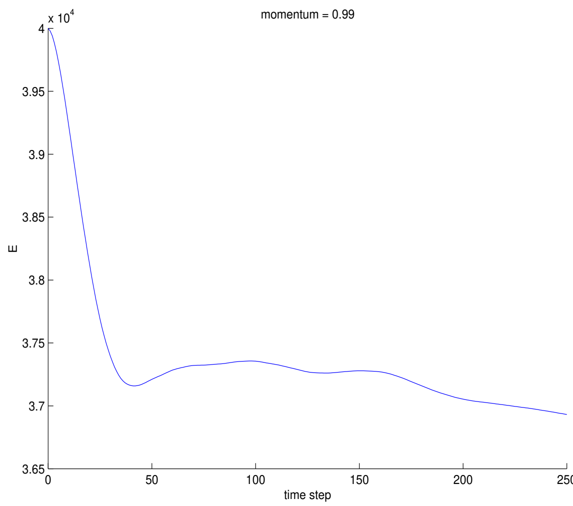


Figure 4: Minimize E with momentum $\eta = 10^{-5}$, $\mu = 0.99$.

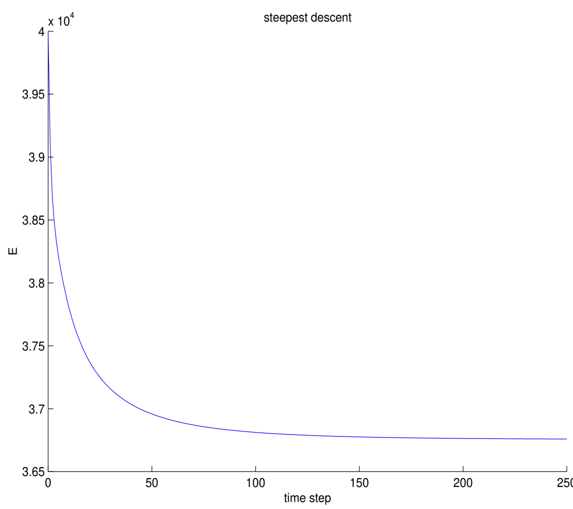


Figure 5: Gradient descent with line search.

3.4 Line Search and Conjugate Gradient

(i) Gradient descent with line search gives E at time step 250 is 3.67591×10^4 (Figure 5), which is smaller than the E's of normal gradient descent and the gradient descent with momentum in 3.3, however bigger than the E of the gradient descent with adaptive learning rate. And, this method need more computation for same time steps. (The time step here refers to the number of calculations of gradient g.)

(ii) Conjugate gradient needs less steps. For $\mathcal{T} = 10$, E at time step 80 is 3.675548×10^4 (Figure 6), which is smaller than all E's above. $\mathcal{T} = 10$ is the best setting among $\mathcal{T} = 2, 10, 100$. When $\mathcal{T} = 100$ and 120 time steps, the E is still worser than $\mathcal{T} = 10$ and only 80 time steps.

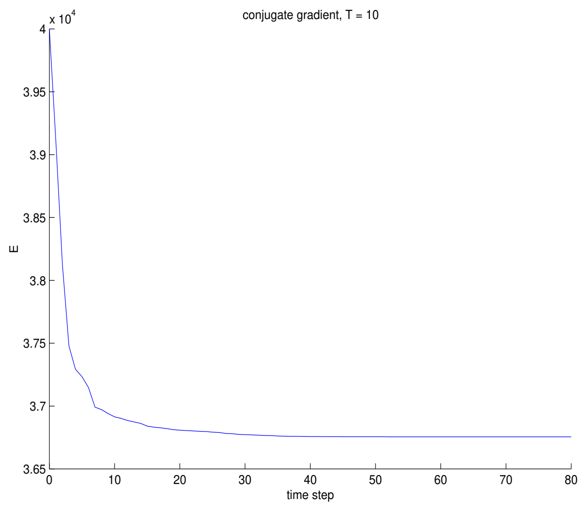


Figure 6: Using conjugate gradient with $\mathcal{T} = 10$.