PC3. Foundations of Machine Learning. (MDC_51006_EP - 2025-2026)

Exercise 1 (Link surrogate loss/prediction) We want to show that the minimizer of

$$\mathbb{E}\left[\bar{\ell}(Y, f(\underline{X}))\right] = \mathbb{E}[l(Y f(\underline{X}))]$$

, where l is a convex non increasing function such that $l(0) = 1$, l is differentiable at 0 and $l'(0) = -1$, is the Bayes classifier $f^\star = \operatorname{sign}(2\mathbb{P}(Y = 1|\underline{X}) - 1)$

1. Write $\mathbb{E}\left[\bar{\ell}(Y, f(\underline{X}))\right]$ as a function $H(f, \eta(\underline{X}))$ where $\eta(\underline{X}) = \mathbb{P}(Y = 1|\underline{X})$.
2. Prove that the optimal $\tilde{f}$ of $H(f, \eta)$ for a given η as the same size than $2\eta - 1$.
3. Conclude

Exercise 2 (Backprop) Let f be a neural network with L hiddern layers parametrized by $W_1, b_1, \ldots, W_L, b_L, W_O, b_O$ by

$$\begin{aligned}
z_1(x) &= W_1 x + b_1\\
h_1(x) &= g_1(z_1(x))\\
&\vdots\\
z_l(x) &= W_l h_{l-1}(x) + b_l\\
h_l(x) &= g_l(z_l(x))\\
&\vdots\\
z_O(x) &= W_O h_L(x) + b_O\\
f(x) &= g_O(z_0(x))
\end{aligned}$$

For the sake of simplicity, we do not denote the dependency on the parameters in the functions. We are nevertheless interested in computing the derivative of

$$F_i = \ell(Y_i, f(X_i))$$

with respect to those parameters.

1. Warmup. Let $u(x) = u_{\text{out}}(u_{\text{cur}}(u_{in}(x, \theta_{\text{in}}), \theta_{\text{cur}}), \theta_{\text{out}})$.
   (a) Verify that

   $$\frac{\partial u^{(d)}}{d\theta_{\text{cur}}^{(d')}}(x) = \sum_k \frac{\partial u_{\text{out}}^{(d)}}{du_{\text{cur}}^{(k)}}(u_{\text{cur}}(u_{\text{in}}(x, \theta_{in}), \theta_{\text{cur}}), \theta_{\text{out}}) \frac{\partial u_{\text{cur}}^{(k)}}{d\theta_{\text{cur}}^{(d')}}(u_{\text{in}}(x, \theta_{\text{in}}), \theta_{\text{cur}})$$

   (b) Using Jacobian matrix notation $\frac{Dv}{dw}$ where $\frac{Dv}{dw}$ is defined by $\left(\frac{Dv}{dw}\right)_{d,'} = \frac{\partial v^d}{dw^{d'}}$, verify that this can be rewritten as

   $$\frac{Du}{d\theta_{\text{cur}}}(x) = \frac{Du_{\text{out}}}{dx_{\text{cur}}}(u_{\text{cur}}(u_{\text{in}}(x, \theta_{in}), \theta_{\text{cur}}), \theta_{\text{out}}) \times \frac{Du_{\text{cur}}}{d\theta_{\text{cur}}}(u_{\text{in}}(x, \theta_{\text{in}}), \theta_{\text{cur}}).$$

2. Using $\theta_l = (\text{flatten}(W_l), b_l)$, where $flatten$ is an operator which transforms a $n \times m$ matrix into a $1 \times (n \times m)$ vector.
   (a) Deduce that

   $$\begin{aligned}
   \frac{DF_i}{d\theta_O} &= \frac{D\ell}{df}(f(X_i)) \times \frac{Df}{dz_O}(z_O(X_i)) \times \frac{Dz_O}{d\theta_O}(h_L(X_i))\\
   \frac{DF_i}{d\theta_l} &= \frac{D\ell}{df}(f(X_i)) \times \frac{Df}{dz_O}(z_O(X_i)) \times \frac{Dz_O}{dh_L}(h_L(X_i)) \times \frac{Dh_L}{dz_L}(z_L(X_i)) \times \frac{Dz_L}{dh_{L-1}}(h_{L-1}(X_i))\\
   &\quad \times \cdots \times \frac{Dh_{l+1}}{dz_{l+1}}(z_{l+1}(X_i)) \times \frac{Dz_{l+1}}{dh_l}(h_l(X_i)) \times \frac{Dh_l}{dz_l}(z_l(X_i)) \times \frac{Dz_l}{d\theta_l}(h_{l-1}(X_i))
   \end{aligned}$$

   with some abuse of notations if $l > L - 1$ and $l = 1$.

(b) Verifiy that

$$\nabla_{\theta_l} F_i = \frac{DF_i}{d\theta_l}^\top$$

(c) Compute

$$\frac{D\ell}{Df}, \quad \frac{Df}{dz_O}, \quad \frac{Dh_l}{dz_l}, \quad \frac{Dz_l}{dh_{l-1}} \quad \text{and} \quad \frac{dz_l}{d\theta_l}$$

3. We are now interested in the complexity of such a computation.
   (a) What are the sizes of those matrices?
   (b) What is the best way to compute the products of those matrices? From left to right? From right to left?
   (c) Why is the left to right direction called backward and the right to left direction called forward?
   (d) Explain why the backward solution is even better when we compute all the derivatives with all the θ_l.