Final Project: Mathematics of Deep Learning (EN 580.745)

Instructor: René Vidal, Biomedical Engineering, Johns Hopkins University

Due Date: 12/15/2018, 11:59PM Eastern Time

Let $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$ be a deterministic input data matrix and let $Y = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N] \in \mathbb{R}^{d \times N}$ be its corresponding output. Let $U \in \mathbb{R}^{d \times r}$ and $V \in \mathbb{R}^{D \times r}$ be, respectively, the output and input weights of a linear neural network with D inputs, d outputs, and a single hidden layer with r neurons. Let $z_k \sim \text{Ber}(\theta_r)$, $\theta_r \in (0, 1)$, be the k-th entry of the random vector $\boldsymbol{z} \in \{0, 1\}^r$ for $k = 1, \ldots, r$. Consider the minimization the following stochastic objective

$$f_r(U, V) = \mathbb{E}_{\boldsymbol{z}} \|Y - \frac{1}{\theta_r} U \text{diag}(\boldsymbol{z}) V^\top X\|_F^2. \tag{1}$$

1. **(15 points)** Derive a stochastic gradient of f_r, i.e., find matrices $\boldsymbol{g}_u(U, V, \boldsymbol{z}) \in \mathbb{R}^{d \times r}$ and $\boldsymbol{g}_v(U, V, \boldsymbol{z}) \in \mathbb{R}^{D \times r}$ such that $\mathbb{E}_{\boldsymbol{z}}[\boldsymbol{g}_u(U, V, \boldsymbol{z})] = \nabla_U f_r(U, V)$ and $\mathbb{E}_{\boldsymbol{z}}[\boldsymbol{g}_v(U, V, \boldsymbol{z})] = \nabla_V f_r(U, V)$. Use $(\boldsymbol{g}_u, \boldsymbol{g}_v)$ to derive an SGD method for minimizing f_r. Explain how SGD relates to the dropout algorithm applied to the squared loss

   $$\ell_r(U, V) = \|Y - UV^\top X\|_F^2. \tag{2}$$

   Write down the GD method for minimizing ℓ_r and compare it with the SGD method for minimizing f_r.

2. **(10 points)** Let $\lambda_r = \frac{1-\theta_r}{\theta_r}$. Show that the stochastic objective f_r is equal to the regularized squared loss

   $$f_r(U, V) = \ell_r(U, V) + \lambda_r \Theta_r(U, V). \tag{3}$$

   Write Θ_r explicitly and explain the effect of the input data X on the regularizer Θ_r.

3. **(25 points)** Assume that $\lambda_r = r\lambda_1 = r\frac{1-\theta_1}{\theta_1}$ and let

   $$\Omega(Z) = \min_{U,V,r} \frac{\lambda_r}{2} \Theta_r(U, V, X) \qquad \text{s.t.} \qquad UV^\top X = Z. \tag{4}$$

   Assume $D \geq N$ and X full column rank. Compute Ω^{**}, i.e., the Fenchel dual of the Fenchel dual of Ω.

4. **(20 points)** Show that if (U, V, r) is a global minimum of $f_r(U, V)$ then $Z = UV^\top$ is a global minimum of:

   $$F(Z) = \frac{1}{2}\|Y - Z\|_F^2 + \Omega^{**}(Z). \tag{5}$$

   Bonus: Write down an algorithm to minimize F. For example, can you compute the proximal operator of Ω^{**}.

5. **(30 points)** Choose `d=30`, `D=100`, `N=50` and `r=10`, and generate `U0=randn(d,r)`, `V0=randn(D,r)`, `X=randn(D,N)` and `Y=U0*V0'*X`.

   (a) Starting at $U_0 = 0$ and $V_0 = 0$, use GD to minimize $\ell_r(U, V)$. Explain how you choose the step size. Plot the loss $\ell_r(U_t, V_t)$ and the error $\epsilon(U_t, V_t) = \|U_t - \texttt{U0}\|_F^2 + \|V_t - \texttt{V0}\|_F^2$ as a function of the number of iterations t. Does the loss converge to 0? Why? Does the error converge to 0? Why?

   (b) Starting at $U_0 = 0$ and $V_0 = 0$, use the dropout algorithm with $\theta_r = 0.5$ to minimize $\ell_r(U, V)$. Plot the loss, regularized loss, and error as a function of the number of iterations. Does the loss converge to 0? Why? Does the error converge to 0? Why? Does the regularized loss converge to 0? Why?

   (c) **Bonus:** Evaluate the proximal operator of $\text{prox}_{\Omega^{**}}(Y)$ and compare with the regularized loss obtained in part (b). Are they the same? Why?

Submission instructions. Please submit a single PDF in Blackboard.