ESE 546, FALL 2020

HOMEWORK 3

SHEIL SARDA [SHEILS@SEAS],
COLLABORATORS: RAHUL M. [RMAG@SEAS]

Solution 1 (Time spent: 5 hours). (1) Prove that co-coercivity implies Lipschitz continuity.

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \Rightarrow \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

According to the Cauchy-Shwarz inequality,

$$\|\langle u, v \rangle\| \leq \|u\| \|v\|$$

Applying Cauchy Schwarz to the RHS of the given inequality and multiplying by L on both sides:

$$L\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \|\nabla f(x) - \nabla f(y)\| \|x - y\| L$$
$$\|\nabla f(x) - \nabla f(y)\|^2 \leq L\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq \|\nabla f(x) - \nabla f(y)\| \|x - y\| L$$

Eliminating the middle term in the inequality above:

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq \|\nabla f(x) - \nabla f(y)\| \|x - y\| L$$
$$\implies \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

$\square$

(2) Prove that the Lipschitz continuity implies co-coercivity. Consider 2 functions:

$$g(z) = f(z) - \langle \nabla f(x), z \rangle$$
$$h(z) = f(z) - \langle \nabla f(y), z \rangle$$

Applying the descent lemma to $g(y)$,

$$\frac{1}{2L} \|\nabla g(y)\| \leq g(y) - g(x)$$
$$\Rightarrow \frac{1}{2L} \|\nabla g(y)\| \leq f(y) - f(x) - \langle \nabla f(x), y \rangle - \langle \nabla f(x), x \rangle$$
$$\Rightarrow \frac{1}{2L} \|\nabla g(y)\| \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$

Therefore,

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|$$
$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|$$

Adding the two above inequalities, we have

$$\langle \nabla f(x) - f(y), x - y \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|$$

(3) Prove that $m \leq \|\nabla^2 f(x)\|_2 \leq L$ Applying the mean value theorem to $\nabla f(x)$,

$$\nabla^2 f(x) = \frac{\nabla f(b) - \nabla f(a)}{b - a}$$

Applying the result from (1) of this question $\implies \nabla^2 f(x) \leq L$.

For any strongly convex function ℓ, the following must hold

$$\ell(w) + \left\langle \nabla \ell(w), w' - w \right\rangle + \frac{m}{2} \left\| w' - w \right\|^2 \leq \ell\left(w'\right)$$

Since we know that $f(x)$ is strongly convex, $\nabla^2 f(x) \succeq m I_{p \times p} = m$ (from Lecture Notes 09). Combining both results, we get:

$$m \leq \|\nabla^2 f(x)\|_2 \leq L$$

$\square$

Solution 2 (Time spent: 6 hours)**.** To prove:

$$\min_{w} \mathop{\mathbb{E}}_{R}\left[\|y-(R\odot X)w\|_2^2\right] = \min_{\tilde{w}} \|y - X\tilde{w}\|_2^2 + \left(\frac{p}{1-p}\right)\tilde{w}^\top \operatorname{diag}\left(X^\top X\right)\tilde{w}$$
$$\text{where } \tilde{w} = (1-p)w$$

For context (from Lecture Notes 07),

- Each row of matrix R consists of the dropout mask for the i^{th} row x^i of the data matrix X.
- Each entry of R is a Bernoulli random variable with probability $1-p$ of being 1.
- For linear regression, dropout is equivalent to weight decay where the coefficient α depends on the diagonal of the data covariance and is different for different weights.
- If a particular data dimension varies a lot $\implies$ X^TX is large, then dropout tries to squeeze its weight to zero.
- If $p = 0$, most activations are retained by the mask and regularization is small.
- Given weights w of a model trained using dropout, we can compute the committee average over models created using dropout masks simply by scaling the weights by a factor $1-p$ $\implies$ $\tilde{w} = (1-p)w$ is the effective weight.

First, we determine $\mathbb{E}[R]$. Each element of R is a Bernoulli random variable. Therefore,

- Case 1: $R_{ij} = 1$. This occurs with probability $1-p$.
- Case 2: $R_{ij} = 0$. This occurs with probability p.

Thus, $\mathbb{E}[R\odot X] = X(1-p)$

We simplify the RHS to eliminate the L2 norm,

$$\min_{w} \mathop{\mathbb{E}}_{R}\left[\|y-(R\odot X)w\|_2^2\right]$$
$$\implies \min_{w} \mathop{\mathbb{E}}_{R}\left[y^2 + w^\top (R\odot X)^\top (R\odot X)w - 2y(R\odot X)w\right]$$
$$\implies \min_{w} y^2 + \mathop{\mathbb{E}}_{R}\left[w^\top (R\odot X)^\top (R\odot X)w\right] - \mathop{\mathbb{E}}_{R}\left[2y(R\odot X)w\right] \qquad \text{Applying LOE}$$
$$\implies \min_{w} y^2 + \mathop{\mathbb{E}}_{R}\left[w^\top (R\odot X)^\top (R\odot X)w\right] - 2yX(1-p)w$$
$$\implies \min_{w} y^2 + \mathop{\mathbb{E}}_{R}\left[w^\top (R\odot X)^\top (R\odot X)w\right] - 2yX\tilde{w}$$
$$\implies \min_{w} y^2 - 2yX\tilde{w} + \tilde{w}^\top X^\top X\tilde{w} - \tilde{w}^\top X^\top X\tilde{w} + \mathop{\mathbb{E}}_{R}\left[w^\top (R\odot X)^\top (R\odot X)w\right] \qquad \text{Completing the square}$$
$$\implies \min_{w} \|y - X\tilde{w}\|_2^2 - \tilde{w}^\top X^\top X\tilde{w} + \mathop{\mathbb{E}}_{R}\left[w^\top (R\odot X)^\top (R\odot X)w\right]$$

The expression $\mathop{\mathbb{E}}_{R}\left[(R\odot X)^\top (R\odot X)\right]$ can be simplified to two cases.

- Case 1: Elements on the main diagonal of $\mathop{\mathbb{E}}_{R}\left[(R_\odot X)^\top (R\odot X)\right] = (1-p) * X^\top X$
- Case 2: Elements off the main diagonal $\mathop{\mathbb{E}}_{R}\left[(R_\odot X)^\top (R\odot X)\right] = (1-p)^2 * X^\top X$

We can express this product in inline notation as

$$
\begin{aligned}
\mathop{\mathbb{E}}_{R}\left[(R \odot X)^\top (R \odot X)\right] &= \operatorname{diag}\left(X^\top X\right)(1-p) + \left(\left(X^\top X\right) - \operatorname{diag}\left(X^\top X\right)\right)(1-p)^2 \\
&= \operatorname{diag}\left(X^\top X\right)(1-p) + \left(X^\top X\right)(1-p)^2 - \operatorname{diag}\left(X^\top X\right)(1-p)^2 \\
&= \operatorname{diag}\left(X^\top X\right)(1-p)p + \left(X^\top X\right)(1-p)^2
\end{aligned}
$$

Substituting this expression into the above equation,

$$
\begin{aligned}
&\min_{w} \|y - X\tilde{w}\|_2^2 - \tilde{w}^\top X^\top X \tilde{w} + \\
&\qquad w^\top \left(\operatorname{diag}\left(X^\top X\right)(1-p)p + \left(X^\top X\right)(1-p)^2\right) w \\
\implies &\min_{w} \|y - X\tilde{w}\|_2^2 - \tilde{w}^\top X^\top X \tilde{w} + w^\top \operatorname{diag}\left(X^\top X\right)(1-p)p\ w + \\
&\qquad w^\top \left(X^\top X\right)(1-p)^2 w \\
\implies &\min_{w} \|y - X\tilde{w}\|_2^2 - \tilde{w}^\top X^\top X \tilde{w} + w^\top \operatorname{diag}\left(X^\top X\right)(1-p)p\ w + \\
&\qquad \tilde{w}^\top \left(X^\top X\right) \tilde{w} \\
\implies &\min_{\tilde{w}} \|y - X\tilde{w}\|_2^2 + \left(\frac{p}{1-p}\right) \tilde{w}^\top \operatorname{diag}\left(X^\top X\right) \tilde{w} \qquad\qquad \text{where } \tilde{w} = (1-p)w
\end{aligned}
$$

$\square$

Solution 3 (Time spent: 6 hours)**.** Population risk of a regression is

$$R(f) = \int |f(x) - y|^2 P(x, y) \mathrm{d}x \mathrm{d}y$$

Prove that model that minimizes the population risk is

$$f^* = \underset{f}{\operatorname{argmin}} R(f) = \mathbb{E}[y \mid x]$$

First simplify the expression inside the integral as

$$\begin{aligned} f(x) - y &= f(x) + \mathbb{E}[y \mid x] - \mathbb{E}[y \mid x] - y \\ &= (y - \mathbb{E}[y \mid x])^2 + 2\,(y - \mathbb{E}[y \mid x])\,(\mathbb{E}[y \mid x] - f(x)) + (\mathbb{E}[y \mid x] - f(x))^2 \end{aligned}$$

Substituting the above into the given integral

$$\begin{aligned} R(f) &= \int |f(x) - y|^2 P(x, y) \mathrm{d}x \mathrm{d}y \\ &= \int |\,(y - \mathbb{E}[y \mid x])^2 + 2\,(y - \mathbb{E}[y \mid x])\,(\mathbb{E}[y \mid x] - f(x)) + (\mathbb{E}[y \mid x] - f(x))^2\,|^2 P(x, y) \mathrm{d}x \mathrm{d}y \\ &= \int (y - \mathbb{E}[y \mid x])^2\, P(x, y) \mathrm{d}x \mathrm{d}y + 2 \int (y - \mathbb{E}[y \mid x])\,(\mathbb{E}[y \mid x] - f(x))\, P(x, y) \mathrm{d}x \mathrm{d}y + \\ &\quad \int (\mathbb{E}[y \mid x] - f(x))^2\, P(x, y) \mathrm{d}x \mathrm{d}y \end{aligned}$$

Since we need to minimize f, ignore the first integral term above as it does not contain f. Next, we simplify the middle term of the integral as follows

$$\begin{aligned} & (y - \mathbb{E}[y \mid x])\,(\mathbb{E}[y \mid x] - f(x)) \\ \implies & (\mathbb{E}[y - \mathbb{E}[y \mid x] \mid x])\,(\mathbb{E}[y \mid x] - f(x)) && \text{Tower property of conditional expectation} \\ \implies & (\mathbb{E}[y \mid x] - \mathbb{E}[\mathbb{E}[y \mid x] \mid x])\,(\mathbb{E}[y \mid x] - f(x)) && \text{Applying LOE} \\ \implies & (\mathbb{E}[y \mid x] - \mathbb{E}[y \mid x])\,(\mathbb{E}[y \mid x] - f(x)) \\ \implies & 0 \end{aligned}$$

Substituting the above result into our original integral, we now need to optimize

$$f^* = \underset{f}{\operatorname{argmin}} \int (\mathbb{E}[y \mid x] - f(x))^2\, P(x, y) \mathrm{d}x \mathrm{d}y$$

Analytically, the above integral will be minimized when $f(x) = \mathbb{E}[y \mid x]$. Thus, the optimal classifier $f^*(x) = \mathbb{E}[y \mid x]$. □

Solution 4 (Time spent: 8 hours)**.** The RNN class built for this question

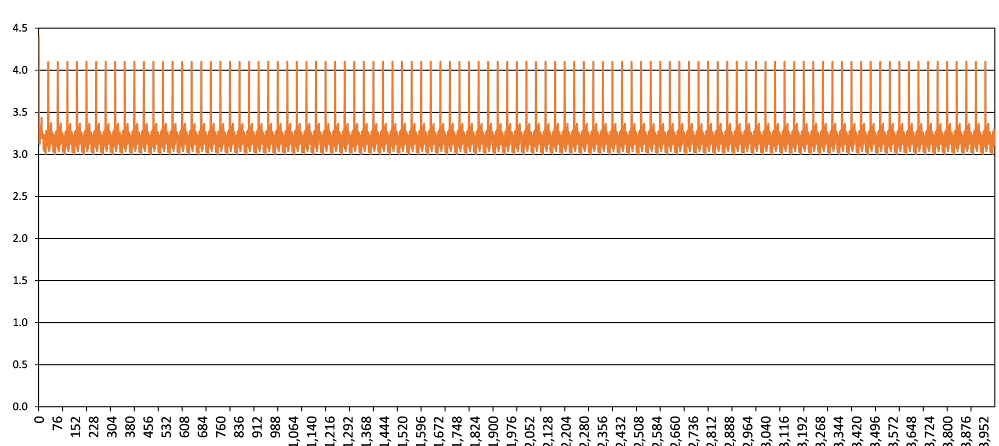
```
class RNN(nn.Module):

  def __init__(self, input_dim, hidden_dim, output_dim,
               no_layers = 1):
    super(RNN, self).__init__()
    self.input_dim = input_dim
    self.hidden_dim = hidden_dim
    self.output_dim = output_dim
    self.no_layers = no_layers

    self.rnn_layer = nn.RNN(self.input_dim,
                            self.hidden_dim, self.no_layers,
                            batch_first = True,
                            nonlinearity='tanh')
    self.linear_out = nn.Linear(self.hidden_dim, self.output_dim)
    self.softmax = nn.LogSoftmax(dim = 1)

  def forward(self, x):
    batch_size = x.size(0)
    hidden = torch.zeros(self.no_layers,
                         batch_size,
                         self.hidden_dim).requires_grad_()
    out, hidden = self.rnn_layer(x, hidden.detach())
    out = out.view(batch_size, len(s))
    return out, hidden
```

Training losses for 100 epochs with 1 Million characters in the training set (40K minibatches), with losses plotted every 1000 minibatches.



Training losses at each epoch.