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:

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

□

(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, $\nabla^2 f(x) \succeq mI_{p \times p} = m$ (from Lecture Notes 09). Combining both results, we get:

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

□

Solution 2 (Time spent: 3 hours)**.** To prove:

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

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.

The RHS can be re-written as:

$$\begin{aligned}
&\min_w \mathbb{E}_R \left[\|y-(R\odot X)w\|_2^2\right] \\
\implies &\min_w \left[\|y-(X(1-p))w\|_2^2\right] && \text{since } R \text{ comprises of Bernoulli variables} \\
\implies &\min_w \left[\|y-X\tilde{w}\|_2^2\right] && \text{by definition of } \tilde{w}
\end{aligned}$$

Now, to go from

$$\min_w \left[\|y-X\tilde{w}\|_2^2\right] \to \min_{\tilde{w}} \|y-X\tilde{w}\|_2^2 + \left(\frac{p}{1-p}\right)\tilde{w}^\top \operatorname{diag}\left(X^\top X\right)\tilde{w}$$

We observe that $\left(\frac{p}{1-p}\right)\tilde{w}^\top \operatorname{diag}\left(X^\top X\right)\tilde{w}$ can be re-written as

$$\begin{aligned}
&\left(\frac{p}{1-p}\right)\tilde{w}^\top \operatorname{diag}\left(X^\top X\right)\tilde{w} \\
\implies &\left(\frac{p}{1-p}\right)\tilde{w}^\top \operatorname{diag}\left(X^\top X\right)w(1-p) \\
\implies &p(1-p)\, w^\top \operatorname{diag}\left(X^\top X\right)w
\end{aligned}$$

Solution 3 (Time spent: 3 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

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

Notes: This is the MSE loss. Minimum of that is the conditional mean of $y|x$. If just trying to optimize $R(f)$, how do you take into account the integral?

The conditional expectation is essentially the integral of $y \times p(y|x) \times dy$. Just looking for the minimum, so write some equations to find the minimum (differentiation with respect to f).

You are finding the min of the regressor by differentiating.

Proof: We can follow