EECS 182 Deep Neural Networks
Fall 2022 Anant Sahai

Discussion 2

1. **Two forms of Ridge Regression** Consider the Ridge Regression estimator,

$$\underset{\mathbf{w}}{\operatorname{argmin}} \|X\mathbf{w} - y\|_2^2 + \lambda\|\mathbf{w}\|^2$$

We know this is solved by

$$\hat{\mathbf{w}} = (X^T X + \lambda I)^{-1} X^T \mathbf{y} \tag{1}$$

An alternate form of the Ridge Regression solution (often called the Kernel Ridge form) is given by

$$\hat{\mathbf{w}} = X^T (XX^T + \lambda I)^{-1} \mathbf{y}. \tag{2}$$

(a) Show that the two solutions for ridge regression are equivalent by algebraic manipulation.

Solution: To show the two forms are equivalent, let's show that

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}$$

With the expression above, let's left-multiply both sides by $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ and right-multiply both sides by $(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})$ to get

$$\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{X}^T$$

And distributing the matrix multiplication we have

$$\mathbf{X}^T\mathbf{X}\mathbf{X}^T + \lambda\mathbf{X}^T = \mathbf{X}^T\mathbf{X}\mathbf{X}^T + \lambda\mathbf{X}^T$$

which we can see is always true as desired.

(b) We know that Ridge Regression can be viewed as finding the MAP estimate when we apply a prior on the (now viewed as random parameters) $\mathbf{W}$. In particular, we can think of the prior for $\mathbf{W}$ as being $\mathcal{N}(\mathbf{0}, I)$ and view the random Y as being generated using $Y = \mathbf{x}^T\mathbf{W} + \sqrt{\lambda}N$ where the noise N is distributed iid (across training samples) as $\mathcal{N}(0,1)$. At the vector level, we have $\mathbf{Y} = X\mathbf{W} + \sqrt{\lambda}\mathbf{N}$, and then we know that when we try to maximize the log likelihood we end up minimizing

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{\lambda}\|X\mathbf{w} - y\|_2^2 + \|\mathbf{w}\|^2 = \underset{\mathbf{w}}{\operatorname{argmin}} \|X\mathbf{w} - y\|_2^2 + \lambda\|\mathbf{w}\|^2.$$

The underlying probability space is that defined by the d iid standard normals that define the $\mathbf{W}$ and the n iid standard normals that give the n different N_i on the training points. Note that the X matrix whose rows consist of the n different inputs for the n different training points are not random.

Based on what we know about joint normality, it is clear that the random Gaussian vectors $\mathbf{W}$ and $\mathbf{Y}$ are jointly normal. Use the following facts to show that the two forms of solution are identical.

- (1) is the MAP estimate for $\mathbf{W}$ given an observation $\mathbf{Y} = \mathbf{y}$.

Solution: This is a fact that's given to us in the problem, but it's not immediately obvious, so some TAs may have re-derived it in their discussion section.
The derivation is given below.
From how we define MAP estimation,

$$MAP(\mathbf{w}|\mathbf{Y}=\mathbf{y}) = \underset{\mathbf{w}}{\text{argmax}}\, f(\mathbf{w}|\mathbf{Y}=\mathbf{y})$$

$$= \underset{\mathbf{w}}{\text{argmax}}\, \frac{f(\mathbf{w},\mathbf{y})}{f(\mathbf{y})}$$

The denominator doesn't affect the argmax since it doesn't depend on $\mathbf{w}$, so we can omit it. Then we can use the chain rule to expand out the numerator

$$= \underset{\mathbf{w}}{\text{argmax}}\, f(\mathbf{w}) f(\mathbf{y}|\mathbf{w})$$

Now, split up the conditional joint density into the product of conditional densities since each element of the $\mathbf{y}$ is independent given $\mathbf{w}$.

$$= \underset{\mathbf{w}}{\text{argmax}}\, f(\mathbf{w}) \prod_{i=1}^{n} f(y_i|\mathbf{w})$$

We can now recall the formula for standard normal pdf is $f_Z(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$. To find $f(y_i|\mathbf{w})$, we know that $y_i = \mathbf{x_i}^T\mathbf{w_i} + \sqrt{\lambda}N_i$, where $N_i \sim \mathcal{N}(0,1)$, so we'd have $\frac{y_i - \mathbf{x_i}^T\mathbf{w}}{\sqrt{\lambda}} \sim \mathcal{N}(0,1)$. Now, plugging in the pdf in the previous expressions we'd have

$$MAP = \underset{\mathbf{w}}{\text{argmax}}\, \frac{e^{-\|\mathbf{w}\|^2/2}}{\sqrt{2\pi}} \prod_{i=1}^{n} \frac{e^{-(\frac{y_i - \mathbf{x_i}^T\mathbf{w_i}}{\sqrt{\lambda}})^2/2}}{\sqrt{2\pi}}$$

We can ignore multiplicative scaling constants (since they don't affect the value of the argmax. Also, since log is a monotonically increasing function, we can take log of both sides without affecting the argmax. Taking logs is useful since it allows us to turn products into sums. So we now have:

$$MAP = \underset{\mathbf{w}}{\text{argmax}} -\frac{\|\mathbf{w}\|^2}{2} - \frac{1}{\lambda}\sum_i \frac{(y_i - \mathbf{x}_i^T\mathbf{w})^2}{2}$$

We can ignore scaling factors again, and arrange all the summation terms in a vector and taking the norm-squared. We can also turn argmax into argmin by negating the objective function:

$$MAP = \underset{\mathbf{w}}{\text{argmin}}\, \|\mathbf{w}\|^2 + \frac{1}{\lambda}\|\mathbf{y} - X\mathbf{w}\|^2$$

Finally, we can multiply through by λ without changing the argmax since it is a positive constant:

$$MAP = \underset{\mathbf{w}}{\text{argmax}}\, \lambda\|\mathbf{w}\|^2 + \|X\mathbf{w} - \mathbf{y}\|^2$$

And now it is the same form as the original Ridge Regression optimization problem.

- For jointly normal random variables, when you condition one set of variables on the values for the others, the resulting conditional distribution is still normal.

- A normal random variable has its density maximized at its mean.
- For jointly normal random vectors that are zero mean, the formula for the conditional expectation is

$$E[\mathbf{W}|\mathbf{Y}=\mathbf{y}] = \Sigma_{WY}\Sigma_{YY}^{-1}\mathbf{y} \tag{3}$$

where the Σ_{YY} is the covariance $E[\mathbf{Y}\mathbf{Y}^T]$ of $\mathbf{Y}$ and $\Sigma_{WY} = E[\mathbf{W}\mathbf{Y}^T]$ is the appropriate cross-covariance of $\mathbf{W}$ and $\mathbf{Y}$.

Solution: We are given that (1) is MAP($\mathbf{w} \mid \mathbf{Y}=\mathbf{y}$). Let's try to find MAP($\mathbf{w} \mid \mathbf{Y}=\mathbf{y}$) in a different way.

We can condition the random variables $\mathbf{W}$ on the values for $\mathbf{Y}=\mathbf{y}$, and we know the resulting conditional distribution $\mathbf{W}|\mathbf{Y}=\mathbf{y}$ is still normal.

Since normal random variables has density maximized at the mean, the MAP estimate is equivalent to the conditional expectation $E[\mathbf{W}|\mathbf{Y}=\mathbf{y}]$. We are given the formula to calculate the conditional expectation.

Before we plug into the formula, let's calculate $\Sigma_{WY} = E[\mathbf{W}\mathbf{Y}^\mathbf{T}]$. Using $\mathbf{Y} = X\mathbf{W} + \mathbf{N}$, we have:

$$\Sigma_{WY} = E[\mathbf{W}(X\mathbf{W} + \sqrt{\lambda}\mathbf{N})^T]$$

Take transposes and distribute, apply linearity of expectation

$$\Sigma_{WY} = E[\mathbf{W}\mathbf{W}^T X^T] + \sqrt{\lambda}E[\mathbf{W}\mathbf{N}^T]$$

Since $\mathbf{W}$ and $\mathbf{N}$ are uncorrelated, and X does not involve any randomness, we can write

$$\Sigma_{WY} = E[\mathbf{W}\mathbf{W}^T]X^T + \sqrt{\lambda}E[\mathbf{W}]E[\mathbf{N}^T]$$

Use the fact that $E[\mathbf{W}\mathbf{W}^T]$ is the covariance matrix of $\mathbf{W}$, which we are given is the identity. Also use the fact that $\mathbf{W}$ and $\mathbf{N}$ are zero-meaned.

$$\Sigma_{WY} = IX^T + 0 = X^T$$

Now let's find $\Sigma_{YY} = E[\mathbf{Y}\mathbf{Y}^T]$.

$$\Sigma_{YY} = E[(X\mathbf{W} + \mathbf{N})(X\mathbf{W} + \mathbf{N})^T]$$

Take transposes and distribute

$$\Sigma_{YY} = E[X\mathbf{W}\mathbf{W}^T X^T] + \sqrt{\lambda}^2 E[\mathbf{N}\mathbf{N}^T] + \sqrt{\lambda}E[X\mathbf{W}\mathbf{N}^T] + \sqrt{\lambda}E[\mathbf{N}\mathbf{W}^T X^T]$$

The cross-terms are 0 since the random vectors are zero-meaned. With some manipulation, we have

$$\Sigma_{YY} = XE[\mathbf{W}\mathbf{W}^T]X^T + \lambda I$$

$$\Sigma_{YY} = XX^T + \lambda I$$

Finally, plugging in Σ_{WY} and Σ_{YY} to the formula for $E[\mathbf{W}|\mathbf{Y}=\mathbf{y}]$, we get $MAP(\mathbf{w}|\mathbf{Y}=\mathbf{y})$ as

$$\hat{\mathbf{w}} = X^T(XX^T + \lambda I)^{-1}\mathbf{y}$$

as desired.

2. Visualizing Backpropagation Consider a simple neural network that takes a scalar real input, has 1 hidden layer with k units in it and a ReLU nonlinearity for those units, and an output linear (affine) layer.

We can algebraically write any function that it represents as

$$y = W^{(2)}(\max(\mathbf{0}, W^{(1)}x + \mathbf{b^{(1)}})) + b^{(2)}$$

Where $x, y \in \mathbb{R}$, $W^{(1)} \in \mathbb{R}^{k\times 1}$, $W^{(2)} \in \mathbb{R}^{1\times k}$, and $\mathbf{b}^{(1)} \in \mathbb{R}^{k\times 1}$, and $b^{(2)} \in \mathbb{R}$. The superscripts are indices, not exponents and the $\max$ given two vector arguments applies the max on corresponding pairs and returns a vector.

For each part, calculate the partial derivative and sketch a small representative plot of the derivative as a function of x. Make sure to clearly label any discontinuities, kinks, and slopes of segments. The subscript i refers to the i-th element of a vector.

(a) $\frac{\partial y}{\partial b^{(2)}}$

Solution:

$$\frac{\partial y}{\partial b^{(2)}} = 1$$

(b) $\frac{\partial y}{\partial w_i^{(2)}}$

Solution:

$$\frac{\partial y}{\partial W_i^{(2)}} = \max(0, W_i^{(1)}x + b_i^{(1)})$$

(c) $\frac{\partial y}{\partial b_i^{(1)}}$

Solution:

$$\frac{\partial y}{\partial b_i^{(1)}} = \begin{cases} W_i^{(2)}, \text{if } W^{(1)}x + b^{(1)} > 0 \\ 0, \text{if } W^{(1)}x + b^{(1)} < 0 \end{cases}$$

(d) $\frac{\partial y}{\partial w_i^{(1)}}$

Solution:

$$\frac{\partial y}{\partial w_i^{(1)}} = \begin{cases} W_i^{(2)}x, \text{if } W^{(1)}x + b^{(1)} > 0 \\ 0, \text{if } W^{(1)}x + b^{(1)} < 0 \end{cases}$$

3. Least Squares and the Min-norm problem from the Perspective of SVD (If time permits)

Consider the equation $X\mathbf{w} = \mathbf{y}$, where $X \in \mathbb{R}^{m\times n}$ is a non-square data matrix, w is a weight vector, and y is vector of labels corresponding to the datapoints in each row of X.

Let's say that $X = U\Sigma V^T$ is the (full) SVD of X. U and V are orthonormal square matrices, and Σ is an $m \times n$ matrix with singular values (σ_i) on the "diagonal".

For this problem, we define $\Sigma^\dagger$ an $n \times m$ matrix with the reciprocals of the singular values ($\frac{1}{\sigma_i}$) along the "diagonal".

(a) First, consider the case where $m > n$, i.e. our data matrix X has more rows than columns (tall matrix) and the system is overdetermined. How do we find the weights w that minimizes the error between $X\mathbf{w}$ and y? In other words, we want to solve $\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2$.

Solution: This is the classic least squares problem. The solution is given by

$$\hat{w} = (X^T X)^{-1} X^T y$$

This can be derived from vector calculus, and also has an elegant interpretation in the context of orthogonal projection of **y** on the column space of X. TAs can go over either during section if students have questions.

(b) Plug in the SVD $X = U\Sigma V^T$ and simplify. Be careful with dimensions!

Solution:

$$(X^T X)^{-1} X^T = (V\Sigma^T U^T U \Sigma V^T)^{-1} V \Sigma^T U^T$$

Since U has orthonormal columns, $U^T U = I$. Notice $\Sigma^T \Sigma$ is a square, $n \times n$ diagonal matrix with squared singular values σ_i^2 along the diagonal.

$$(X^T X)^{-1} X^T = (V\Sigma^T \Sigma V^T)^{-1} V \Sigma^T U^T$$

Apply the fact that $(AB)^{-1} = B^{-1}A^{-1}$, and that $V^{-1} = V^T$ since the matrix is orthonormal.

$$(X^T X)^{-1} X^T = V(\Sigma^T \Sigma)^{-1} V^T V \Sigma^T U^T$$

Simplify since $V^T V = I$.

$$(X^T X)^{-1} X^T = V(\Sigma^T \Sigma)^{-1} \Sigma^T U^T$$

Notice that $(\Sigma^T \Sigma)^{-1}\Sigma^T$ is an $n \times m$ matrix with the reciprocals of the singular values, $\frac{1}{\sigma_i}$, on the "diagonal". We can call this matrix $\Sigma^\dagger$. Note that this isn't a true matrix inverse (since the matrix Σ is not square). So we can write our answer as

$$(X^T X)^{-1} X^T = V \Sigma^\dagger U^T$$

You should draw out the matrix shapes and convince yourself that all the matrix multiplications make sense.

(c) What happens if we left-multiply X by our least squares solution?

Solution: $(X^T X)^{-1} X^T X = I$. We can also see this from our SVD interpretation,

$$V\Sigma^\dagger U^T U \Sigma V^T = V \Sigma^\dagger \Sigma V^T = V V^T = I$$

Students should make sure to understand why $\Sigma^\dagger \Sigma = I$ (What are the dimensions, and what are the entries?)

This is why the least-squares solution is called the left-inverse.

(d) Now, let's consider the case where $m < n$, i.e. the data matrix X has more columns than rows and the system is underdetermined. There exist infinitely many solutions for w, but we seek the minimum-norm solution, ie. we want to solve $\min \|\mathbf{w}\|^2 \text{s.t.} X\mathbf{w} = \mathbf{y}$. What is the minimum norm solution?

Solution: The min-norm problem is solved by

$$\mathbf{w} = X^T (X X^T)^{-1} \mathbf{y}$$

(We can see this by choosing $\mathbf{w}$ that has a zero component in the nullspace of X, and thus $\mathbf{w}$ is in the range of X^T. TAs can go through the derivation if students have questions).

(e) Plug in the SVD $X = U\Sigma V^T$ and simplify. Be careful with dimensions!

Solution:

$$\begin{aligned} X^T(XX^T)^{-1} &= (U\Sigma V^T)^T(U\Sigma V^T(U\Sigma V^T)^T)^{-1} \\ &= V\Sigma^T U^T(U\Sigma V^T V\Sigma^T U^T)^{-1} \\ &= V\Sigma^T U^T U(\Sigma\Sigma^T)^{-1}U^T \\ &= V\Sigma^T(\Sigma\Sigma^T)^{-1}U^T \end{aligned}$$

Here, we have that $\Sigma^T(\Sigma\Sigma^T)^{-1}$ is an $n \times m$ matrix with the reciprocals of the singular values, $\frac{1}{\sigma_i}$, on the "diagonal". We can call this matrix $\Sigma^\dagger$ so that we have

$$= V\Sigma^\dagger U^T$$

(f) What happens if we right-multiply X by our min-norm solution?

Solution: Similar to the previous part, $XX^T(XX^T)^{-1} = I$. This can also be seen from the SVD perspective.

This is why the min-norm solution is called the right-inverse.

