QUADRATIC LOSS: A WORKED EXAMPLE

From last class:

$$\ell_2(y) = (1-y)^2 = (y-1)^2 \qquad m(x,y,\omega) = \langle x,\omega\rangle y \qquad R_2(\omega) = \lambda\|\omega\|^2$$

$$D\ell_2(y) = 2(y-1) \qquad D_\omega m(x,y,\omega) = xy \qquad D_\omega R_2(\omega) = 2\lambda\omega$$ Derivatives

$$L(\omega) = \left(\sum_{(x,y)} \ell_2 \circ m(x,y,\omega)\right) + R_2(\omega)$$

$$D_\omega L(\omega) = \left(\sum_{(x,y)} D\ell_2 \circ m(x,y,\omega) \times D_\omega\, m(x,y,\omega)\right) + D_\omega R_2(\omega)$$

$$= \left(\sum_{(x,y)} 2(\langle x,\omega\rangle y - 1)\cdot xy\right) + 2\lambda\omega$$

Find the extreme value (set gradient to zero):

$$\left(\sum_{(x,y)} (\langle x,\omega\rangle - 1)xy\right) + \lambda\omega = 0$$

$$\left(\sum \langle x,\omega\rangle xy\right) + \lambda\omega = \sum xy$$

Now rearrange so we can write left side as $M\omega$:

$$\sum y^2 xx^T\omega + \lambda\omega = \sum xy$$

$$\left(\left(\sum y^2 xx^T\right) + \lambda I\right)\omega = \sum xy$$

$$(A + \lambda I)\omega = b$$

$$M\omega = b$$

$$\omega = M^{-1}b$$

But we can do better! Determining best λ is cheap:

$$A = Q \Lambda Q^T$$
$$A + \lambda I = Q \Lambda Q^T + Q(\lambda I) Q^T$$
$$M = Q(\Lambda + \lambda I) Q^T$$
$$M^{-1} = Q(\Lambda + \lambda I)^{-1} Q^T$$

↑ This is trivial to compute!

Behavior is intimately linked to eigenvalues and eigenvectors of $\underline{\underline{A}}$.

GRADIENT DESCENT

Let's write our loss in the format of the analysis in Goh's paper:

$$
\begin{aligned}
L(w) &= \sum_{x,y} (l_2 \circ m)(x,y,m) + R_2(w) \\
&= \sum_{x,y} \left(\langle x,w\rangle_y - 1\right)^2 + \lambda \|w\|^2 \\
&= \sum_{x,y} \left(x^T w y - 1\right)^2 + \lambda w^T w \\
&= \sum_{x,y} \left(w^T x x^T w y^2 - 2 x^T w y + 1\right) + \lambda w^T w \\
&= w^T \left(\sum_{x,y} x x^T y^2\right) w - 2\left(\sum_{x,y} x^T y\right) w + \sum_{x,y} 1 + \lambda w^T w \\
&= w^T \left(\left(\sum x x^T y^2\right) + \lambda I\right) w - \left(2 \sum x^T y\right) w + \sum 1
\end{aligned}
$$

$$
\begin{aligned}
f(w) &= \frac{1}{2} w^T A w - b^T w \\
\nabla_w f(w) &= \frac{2}{2} A w - b
\end{aligned}
$$

$\nabla_w f = 0$ (we will use this soon)

$A w^* = b$

$w^* = A^{-1} b$

GRADIENT ITERATION

$$w^{k+1} = w^k - \alpha(Aw^k - b)$$

A is symmetric, so admits eigen analysis:

$$A = Q \Lambda Q^T$$

Now let's look at consecutive iterations:

$$w^{k+1} = w^k - \alpha(Aw^k - b)$$
$$w^{k+1} = (I - \alpha A) w^k + \alpha b$$

$$I - \alpha A = QQ^T - \alpha Q \Lambda Q^T$$
$$= Q(I - \alpha \Lambda) Q^T$$

$$(I - \alpha A)^2 = Q(I - \alpha \Lambda) Q^T Q (I - \alpha \Lambda) Q^T$$
$$= Q(I - \alpha \Lambda)^2 Q^T$$

$I - \alpha \Lambda$ is a simpler expression than $I - \alpha A$: the former is a diagonal matrix. Let's try to get gradient descent to look like $I - \alpha \Lambda$. First, let's translate the whole problem so that the minimum is at zero, and then let's write everything in terms of vectors multiplied by Q^T:

$$\omega^* = A^{-1}b$$

$$\omega^* = Q\Lambda^{-1}Q^T b$$

$$Q^T\omega^* = \Lambda^{-1}Q^T b$$

$$\Lambda Q^T\omega^* = Q^T b$$

$$x^k = Q^T(\omega^k - \omega^*)$$

$$Qx^k = \omega^k - \omega^*$$

$$\omega^k = Qx^k + \omega^*$$

$$\omega^{k+1} = \omega^k - \alpha(A\omega^k - b)$$

$$Q^T\omega^{k+1} = Q^T\omega^k - \alpha Q^T(A\omega^k - b)$$

$$Q^T\omega^{k+1} - Q^T\omega^* = Q^T\omega^k - Q^T\omega^* - \alpha Q^T(A\omega^k - b)$$

$$x^{k+1} = x^k - \alpha(Q^TQ\Lambda Q^T\omega^k - Q^Tb)$$

$$x^{k+1} = x^k - \alpha(\Lambda Q^T\omega^k - \Lambda Q^T\omega^* + \Lambda Q^T\omega^* - Q^Tb)$$

$$x^{k+1} = x^k - \alpha\Lambda x^k$$

$$x^{k+1} = (I - \alpha\Lambda)x^k$$

$$x^k = (I - \alpha\Lambda)^k x^0$$

$$Qx^k = Q(I - \alpha\Lambda)^k x^0$$

$$Qx^k + \omega^* = Q(I - \alpha\Lambda)^k x^0 + \omega^*$$

$$\omega^k = Q(I - \alpha\Lambda)^k x^0 + \omega^*$$

$$\omega^k = Q(I - \alpha\Lambda)^k Q^T(\omega^k - \omega^*) + \omega^*$$

How do we make sure that ω^k will go to ω^*?

The error $\omega^k - \omega^*$ must go to zero. In addition, each coordinate of $Q^T(\omega^k - \omega^*)$ goes to zero (or fails to) independently of the other: each gets scaled by $(I - \alpha\Lambda)$ every iteration.

This is the true picture of gradient descent: the error vector is multidimensional, and the amount it shrinks on coordinate i (of the eigenspace) is given by $|1 - \alpha \lambda_i|$, and each coordinate shrinks independently of the other.

Given an initial guess x^0, the magnitude of the (quadratic) error is:

$$\sum (x_i^0)^2$$

Even if each x_i is identical, each iteration reduces each coordinate proportionally to λ_i, so the features corresponding to smaller eigenvalues will take longer to converge to their minimum. The difference between early iterations and late iterations, then, is due to features coming from small eigenvalues.

WHAT ABOUT REGULARIZATION?

Intuitively, regularization should make optimization want smaller w vectors,

$$L(w) = \sum_{x,y} \text{loss-of-sample}\left(y, \text{margin}(x, y, w)\right) + \frac{\lambda}{2}\|w\|^2$$

This becomes directly obvious when we consider a gradient descent update:

$$\nabla_w L(w) = \left(\nabla_w \sum \text{loss-of-sample}\right) + \lambda w$$

So a step taken with a regularized loss is like first moving a little towards zero, and then taking a regular step:

$$\begin{aligned} w^{k+1} &= w^k - \alpha \nabla_w L(w) \\ &= w^k - \alpha\left(\left(\nabla_w \sum \text{loss-of-sample}\right) + \lambda w^k\right) \\ &= w^k - \alpha\left(\nabla_w \sum \text{loss-of-sample}\right) - \alpha\lambda w^k \\ &= (1 - \alpha\lambda) w^k - \alpha\left(\nabla_w \sum \text{loss-of-sample}\right) \end{aligned}$$

(The notation here gets a little confusing: λ is the regularization term, and λ_i are the eigenvalues of A)

Finally, let's consider the difference between the solution of the regularized problem and the solution of the unregularized problem. The difference, incredibly, is analogous to that of choosing whether to stop gradient descent early!

$$2\ f(\omega) = \omega^T A\omega - 2\omega^T b + \lambda\|\omega\|^2$$

$$2\ \nabla f = 2A\omega - 2b + 2\lambda\omega$$

$$\nabla f = (A+\lambda I)\omega - b$$

$$(A+\lambda I)\omega = b$$

$$\omega^* = A^{-1}\ b$$

$$\omega'^* = (A+\lambda I)^{-1} b$$

$$\omega^* - \omega'^* = (A^{-1} - (A+\lambda I)^{-1})b$$

But recall our eigenanalysis:

$$A = Q\Lambda Q^T$$

$$I = QQ^T$$

$$\omega^* - \omega'^* = ((Q\Lambda Q^T)^{-1} - (Q\Lambda Q^T + \lambda QQ^T)^{-1})b$$

$$= ((Q\Lambda Q^T)^{-1} - (Q(\Lambda+\lambda I)Q^T)^{-1})b$$

Stupid linear algebra tricks, now:

1) If Q is orthogonal, $Q^{-1} = Q^T$
2) $(AB)^{-1} = B^{-1}A^{-1}$

$$(Q\Lambda Q^T)^{-1} = (Q^T)^{-1}(Q\Lambda)^{-1}$$
$$= Q\Lambda^{-1}Q^{-1}$$
$$= Q\Lambda^{-1}Q^T \qquad (!)$$

$$w^* - w'^* = \left(Q\,\mathrm{diag}\left(\frac{1}{\lambda_i}\right)Q^T - Q\,\mathrm{diag}\left(\frac{1}{\lambda+\lambda_i}\right)Q^T\right)b$$
$$= Q\,\mathrm{diag}\left(\frac{\lambda}{\lambda+\lambda_i}\cdot\frac{1}{\lambda_i}\right)Q^T b \qquad \left(\frac{1}{\lambda_i} - \frac{1}{\lambda+\lambda_i}\right)$$

What happens when $\lambda \gg \lambda_i$?

What happens when $\lambda_i \gg \lambda$?