Week 7: Friday, Oct 5

Logistics

1. The take home MT exam will be posted next Wednesday, Oct 9, and is due on Monday, Oct 22 (in lecture or via CMS). The exam will be open book.
2. This will be the last lecture on least squares problems. We will turn to eigenvalue problems on Wednesday.

The need for regularization

Suppose $Z \in \mathbb{R}^n$ is a vector whose components are independent standard normal random variables. Then $W = AZ$ has a multivariate normal distribution with covariance AA^T. Now, consider a least squares problem with noise:

$$\text{minimize } \|AX - B\|_2^2$$

where $B = b + \epsilon Z \in \mathbb{R}^n$ is the true right hand side b contaminated by independent normal noise with variance ϵ^2 in each component. Writing the pseudoinverse in terms of the economy SVD as $A^\dagger = V\Sigma^{-1}U^T$, we have that X is a multivariate normal random variable with mean $x = A^\dagger b$ and covariance $\epsilon^2 V\Sigma^{-2}V^T$. Even more convenient, consider $\tilde{X} = V^T X$, which has mean $V^T x$ and covariance $\epsilon^2\Sigma^{-2}$.

The random variable $\tilde{X}$ is the best linear unbiased estimator for $\tilde{x}$; but what about the variance of this estimator? If some of the singular values of A are very small relative to ϵ, the variance of the corresponding components of $\tilde{X}$ may be huge, even though the expected value of $\tilde{X}$ is correct. If we have reason that $\tilde{x}_i$ is modest but $\sigma_i \ll \epsilon$, we may be better off estimating $\tilde{x}_i$ by zero instead of using $\tilde{X}_i$. That is, we trade in the unbiased estimator $(\tilde{X}_i)$ with a biased estimator (0) which has better variance. If we do this for each component $i > r$, we are left with the new estimator

$$\hat{X} = V_r \Sigma_r^{-1} U_r^T,$$

where V_r and U_r consist of the first r columns of V and U, and $\Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$. This is sometimes called the *truncated SVD solution* to the least squares problem.

The truncated SVD solution is one example of a *regularized* solution to an ill-posed least squares problem, and in the previous lecture we briefly described another regularization approach based on using a truncated pivoted QR decomposition for factor selection. Informally, ill-posed problems are ones in which A is poorly conditioned, so that small errors in the right hand side lead to giant coefficients in the solution vector. The problem is that the data simply does not provide enough information to yield a stable, accurate estimate of the solution vector. So we make a trade, seeking instead an approximate solution vector that fits the data nearly as well as the full least squares solution, but without the potentially wild coefficients.

Filtering and Tikhonov regularization

We can write the regularized least squares solution $\tilde{x}$ obtained by a truncated SVD as

$$\tilde{x} = V\tilde{\Sigma}^{-1}U^T b$$

where $\tilde{\sigma}_i^{-1} = f(\sigma_i)\sigma_i^{-1}$, and the filter function f is a simple cutoff function:

$$f(\sigma) = \begin{cases} 1, & \sigma \geq \sigma_r \\ 0, & \text{otherwise.} \end{cases}$$

Unfortunately, computing the SVD is an expensive operation. What if we could instead use a different filter that still discards – or at least mitigates – the influence of the small singular values? One such filter function is the *Tikhonov filter*:

$$f(\sigma) = \frac{\sigma^2}{\sigma^2 + \alpha^2}.$$

The filtered inverse singular values are

$$\tilde{\sigma}_i^{-1} = \frac{\sigma_i}{\sigma_i^2 + \alpha^2}$$

For values of σ that are large relative to α, $f(\sigma)$ is close to one, and $\tilde{\sigma}_i^{-1} \approx \sigma_i^{-1}$. But for values of σ close to zero, $f(\sigma)$ approaches zero like σ^2, which is a fast enough decay rate to “damp out” the influence of small singular values.

The solution to the Tikhonov-regularized least squares problem can be written as

$$\begin{aligned} \tilde{x} &= V(\Sigma^2 + \alpha^2 I)^{-1} \Sigma U^T b \\ &= (V \Sigma^2 V^T + \alpha^2 I)^{-1} V \Sigma U^T b \\ &= (A^T A + \alpha^2 I)^{-1} A^T b \end{aligned}$$

That is, we can apply the filtering operation without using the SVD at all! If we look more closely, we might note that $\tilde{x}$ also satisfies the normal equation for the problem

$$\text{minimize } \left\| \begin{bmatrix} A \\ \alpha I \end{bmatrix} \tilde{x} - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|^2 = \|A\tilde{x} - b\|^2 + \alpha^2 \|\tilde{x}\|^2.$$

Looking at this formulation, we can see again the basic tradeoff of regularization: find a solution which is consistent with the data (the first term), but penalize solutions if they are inconsistent with prior assumptions about things like the size of the coefficients (the second term).