Tikhonov-Regularized Linear Problem

Kaibo Tang

August 28, 2024

1 Introduction

The purpose of this note is to give some intuitions about how Tikhonov regularization helps iterative algorithms converge to a plausible solution. In particular, I will provide intuitions from four different perspectives–two from the standpoint of linear algebra, one front a geometric perspective, and one statistical interpretation.

2 Problem Statement

Consider the linear systems of form

$$Ax = b, \tag{1}$$

where we assume A to have shape $m \times n$, x to have shape $n \times 1$, and b to have shape $m \times 1$.

When $m > n$, the problem has infinite many solutions. In this case, we say the system is under-determined and we say matrix A is ill-conditioned. When $m = n$, the problem has a unique solution when A is non-singular, and has a infinite many solutions otherwise. When $m < n$, the problem usually does not have a solution. However, we can find a $\hat{x}$ such that it minimizes the squared ℓ_2-norm of the residual, i.e.

$$\hat{x} = \arg\min_{x} \|Ax - b\|_2^2. \tag{2}$$

For the sake of simplicity, in this note, I will focus solely on the case where $m > n$ where the system is under-determined. Since the system has infinite many solutions, we might be interested in finding the solution whose ℓ_2-norm is the smallest. This motivates the use of Tikhonov regularization, which does so by solving the minimization problem of form

$$\hat{x} = \arg\min_x \|Ax - b\|_2^2 + \lambda \|x\|_2^2. \tag{3}$$

To find $\hat{x}$ as in (3), we differentiate the function we wish to minimize and set the gradient to 0, i.e.,

$$0 = \frac{d}{dx}\left[\|Ax - b\|_2^2 + \lambda \|x\|_2^2\right] = 2(Ax - b)^T A + 2\lambda x^T. \tag{4}$$

Simplification of (4) yields the normal equation

$$(A^T A + \lambda I)x = A^T b. \tag{5}$$

At this point, we almost arrive at our first intuition, which we will discuss without further ado.

3 Intuition 1: Positive Definite Matrix

Equation (5) suggests that, $\forall \lambda > 0$, we have a closed form solution given by

$$\hat{x} = (A^T A + \lambda I)^{-1} A^T b, \tag{6}$$

where the square matrix $A^T A + \lambda I$ is positive definite and is thus invertible. To see why $A^T A + \lambda I$ is positive definite, notice that $\forall x \neq 0$, we have

$$x^T (A^T A + \lambda I)x = \|Ax\|_2^2 + \lambda \|x\|_2^2 > 0. \tag{7}$$

4 Intuition 2: Condition Number

If any arbitrarily small $\lambda > 0$ makes $A^T A + \lambda I$ invertible, why bother picking a bigger λ? The second intuition explains the benefit of picking a bigger λ.

Recall condition number from a typical undergraduate-level numerical analysis course. In our case, the condition number $\kappa(A)$ quantifies the sensitivity of x to slight perturbations in b. A smaller $\kappa(A)$ would suggest that the solution $\hat{x}$ to the system in (1) does not change much when slight perturbations, e.g., noise, is applied to b.

In the following figure, we demonstrate the effect of λ on the condition number of $A^T A + \lambda I$, where $A = \begin{bmatrix} 1 & 1 \end{bmatrix}$.

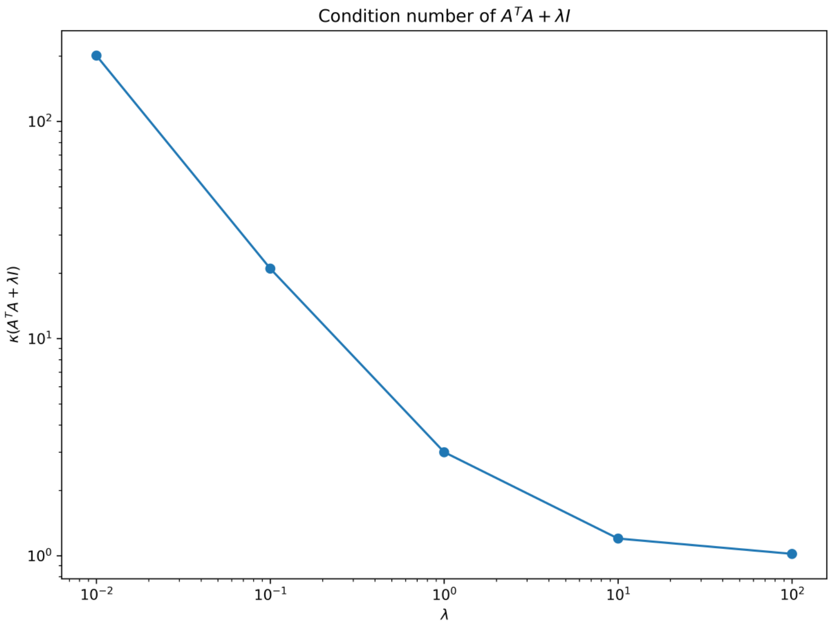

Figure 1: Condition number of $A^T A + \lambda I$ for $\lambda \in \{0.01, 0.1, 1, 10, 100\}$.

5 Intuition 3: Loss Landscape

For my visual learners out there, I visualize the loss landscape for different λ. In particular, consider the same linear system with $A = \begin{bmatrix} 1 & 1 \end{bmatrix}$ and $b = 1$. Notice that the solution to the system is the set of all points on the line $x_1 + x_2 = 1$. In particular, the solution with the least ℓ_2-norm is $(0.5, 0.5)$. The loss function is given by

$$f(x) = \|Ax - b\|_2^2 + \lambda \|x\|_2^2 = (x_1 + x_2 - 1)^2 + \lambda(x_1^2 + x_2^2). \tag{8}$$

From the figure, observe that the unique solution to the Tikhonov-regularized linear problem can get arbitrarily close to $(0.5, 0.5)$, the solution with the least ℓ_2-norm when we pick λ sufficiently small. However, the corresponding loss landscape does not look too good–once we get to the valley, i.e., close to the line $x_1 + x_2 = 1$, the gradient becomes too small. On the other hand, when we have a large λ, the loss landscape looks great. But since the optimization problem now is dominated by the regularization term, the unique solution to the Tikhonov-regularized linear problem is close to the origin, whose ℓ_2-norm is close to 0.

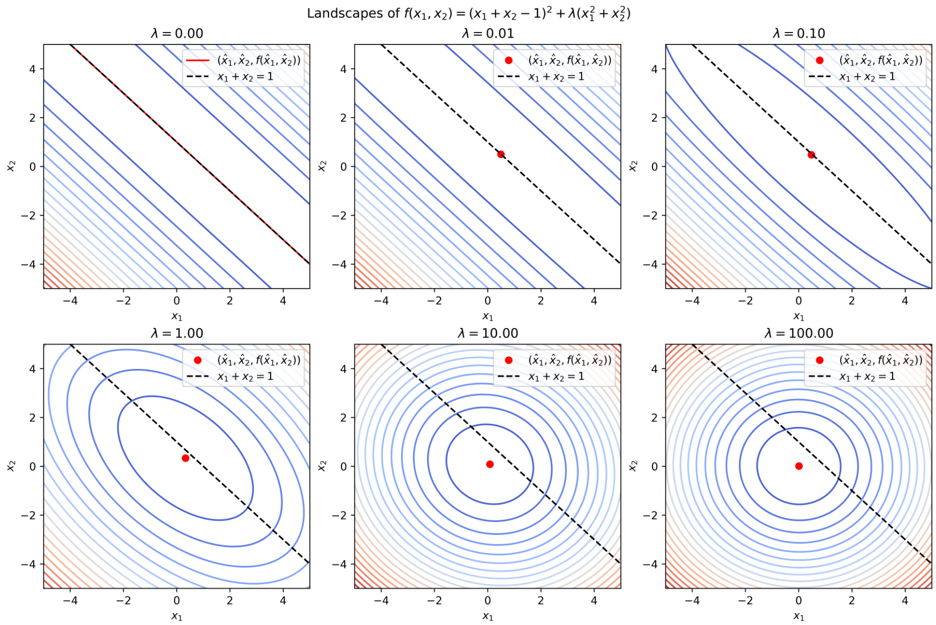

Figure 2: Loss landscape (contour map) for $\lambda \in \{0, 0.01, 0.1, 1, 10, 100\}$.

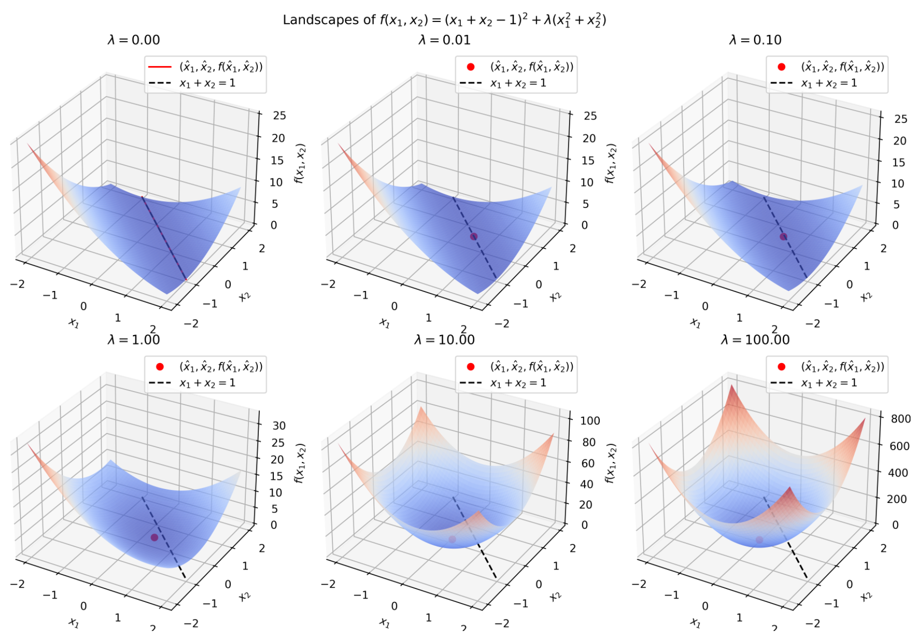

Figure 3: Loss landscape (surface) for $\lambda \in \{0, 0.01, 0.1, 1, 10, 100\}$.

6 Intuition 4: Statistical Interpretation

Since I am a biostatistics major myself, I would like to finish the note with a statistical interpretation.

We first recall the system of linear equations in (1),

$$Ax = b. \tag{9}$$

To motivate the following interpretation, we assume that we wish to recover the underlying signal x from noisy observation b corrupted by additive white Gaussian noise, i.e.,

$$b = Ax + \epsilon \quad \text{where } \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I). \tag{10}$$

Note how the independence and homoscedasticity assumptions are implied here.

The maximum a posteriori (MAP) estimate of x from b is given by

$$\hat{x} = \arg\max_x p(x|b) = \arg\min_x [-\ln p(b|x) - \ln p(x)], \tag{11}$$

where the second equality came from Bayes rule. To simplify the log-likelihood term in (11), notice that

$$-\ln p(b|x) = -\ln (2\pi)^{-\frac{n}{2}} |\sigma_\epsilon^2 I|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(b - Ax)^T (\sigma_\epsilon^2 I)^{-1} (b - Ax)\right] \tag{12}$$

$$= \frac{1}{2\sigma_\epsilon^2} \|b - Ax\|_2^2 + \text{const.} \tag{13}$$

Now, we are left with the log-prior term. The prior term, as is suggested by the name, should carry some form of "prior knowledge" regarding the distribution of x. In medical imaging, this "prior knowledge" often comes in the form of sparsity, e.g. sparsity in the Fourier-transformed frequency domain, wavelet-transformed coefficient domain, spherical harmonics coefficient domain, etc., which are all beyond the scope of this note. Here, we make a very simple assumption about what we know about x in terms of "prior knowledge", i.e., that

$$x \sim \mathcal{N}(0, \sigma_x^2 I). \tag{14}$$

However, typically, the signal x we are trying to estimate, e.g., an image, or a time-series, often display some form of auto-correlation and the assumption in (14) is rarely satisfied but here we go. With the assumption in (14), we can now simplify the log-prior term in (11). Notice that, similar to (12) and (13), we have

$$-\ln p(x) = \frac{1}{2\sigma_x^2}\|x\|_2^2 + \text{const.} \tag{15}$$

Now, the MAP estimate of x is given by

$$\hat{x} = \arg\min_x \left[\frac{1}{2\sigma_\epsilon^2}\|Ax - b\|_2^2 + \frac{1}{2\sigma_x^2}\|x\|_2^2\right]. \tag{16}$$

Observe that without the log-prior term, the MAP estimator agrees with the ordinary least square (OLS) estimator. Further simplify (16), we have

$$\hat{x} = \arg\min_x \left[\|Ax - b\|_2^2 + \frac{\sigma_\epsilon^2}{\sigma_x^2}\|x\|_2^2\right]. \tag{17}$$

I would encourage the reader to always consider from a MAP point-of-view before attempting to pick an appropriate λ for Tikhonov regularization. For example, a low signal-to-noise ratio (SNR) during the acquisition process of b would suggest a higher σ_ϵ^2, in which case a larger λ should be picked accordingly. Alternatively, if we know enough about the real distribution of x and are confident enough that the underlying x indeed has small ℓ_2-norms, we can also pick a large λ.