CISC 371 Class 26

Constrained Least Squares And Tikhonov Regularization

Texts: [1] pp. 218–222,261–266; [2] pp. 305–317

Main Concepts:

- *Ordinary least squares: the normal equation*
- *Constrained least squares: restriction θ on OLS weight vector*
- *Tikhonov regularization*
- *Total variation*
- *Denoising using total variation*
- *Denoising as Tikhonov regularization*

Sample Problem, Machine Inference: How can we optimize a convex objective that is constrained by a single scalar hyper-parameter?

In this course, part of the intellectual landscape that we have explored has included the minimization of an objective function $f(\vec{w})$ that is subject to an equality constraint $p(\vec{w}) = 0$. We observed that, if this problem has a solution, then there is a Lagrange function that has a stationary point at the solution.

We will explore simple constraints on a least-squares regression. A particularly simple formulation is called constrained least squares; a more general formulation is called regularization, which we will examine as a particular formulation that is called Tikhonov regularization.

26.1 Ordinary Least Squares

Ordinary least squares (OLS) is the conventional name in the area of optimization for a problem that is familiar from prerequisite material, where is is often called "linear least squares". The more general approximation problem is: given a set of m independent data vectors $\vec{x}_i$, and m dependent data readings y_i, to find a weight vector $\vec{w}$ that approximates the dependent data values as

$$\vec{x}_i \cdot \vec{w} \approx y_i$$

We can "vectorize" this approximation by gathering the independent data vectors $\vec{x}_i$ into a design matrix X, and gathering the dependent data readings y_i into a data vector $\vec{y}$. When we create the design matrix X, we have two choices: a data vector $\vec{x}_i$ can be the i^{th} column of X, or the transpose of the data vector $\vec{x}_i^T$ can be the i^{th} row of X. We will follow the convention in statistics that uses the second version, so we will define the design matrix as

$$X_{m\times n} \stackrel{\text{def}}{=} \begin{bmatrix} \vec{x}_1^T \\ \vec{x}_2^T \\ \vdots \\ \vec{x}_m^T \end{bmatrix} \tag{26.1}$$

Using Equation 26.1, the problem is to find a weight vector $\vec{w}$ so that

$$X\vec{w} \approx \vec{y} \tag{26.2}$$

We will assume that the data vectors $\vec{x}_i$ either exactly determine or over-determine the weight vector $\vec{w}$. Mathematically, we will assume that the design matrix X is full rank, so rank$(X) = n$.

In the first part of this class, we will explore the solution of Equation 26.2 as an unconstrained optimization problem. In the second part, we will impose a constraint on the weight vector $\vec{w}$. This constraint will be managed by forming a Lagrange equation and we will consider two specific types of solutions to the problem.

We will simplify the approximation problem of Equation 26.2 by restricting the approximation. The vector of residual errors of the approximation is the difference between the approximated values $X\vec{w}$ and the dependent data readings $\vec{y}$, so

$$\vec{r}(\vec{w}) \stackrel{\text{def}}{=} X\vec{w} - \vec{y} \tag{26.3}$$

The OLS problem is to find the weight vector $\vec{w}^*$ that minimizes the sum of squares of the residual errors that are defined in Equation 26.3, which is

$$\begin{aligned} \vec{w}^* &= \underset{\vec{w}\in\mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{m} (r_i(\vec{w}))^2 \\ &= \underset{\vec{w}\in\mathbb{R}^n}{\operatorname{argmin}} \|\vec{r}(\vec{w})\|^2 \\ &= \underset{\vec{w}\in\mathbb{R}^n}{\operatorname{argmin}} \|X\vec{w} - \vec{y}\|^2 \end{aligned} \tag{26.4}$$

We can formulate the OLS problem in Equation 26.4 by setting an objective function $f(\vec{w})$ to be

$$\begin{aligned} f(\vec{w}) &= \|X\vec{w} - \vec{y}\|^2 \\ &= [X\vec{w} - \vec{y}]^T[X\vec{w} - \vec{y}] \\ &= \vec{w}^T[X^TX]\vec{w} + [-2\vec{y}^TX]\vec{w} \end{aligned} \tag{26.5}$$

In Equation 26.5, we have dropped the $\vec{y}^T\vec{y}$ term because the term is constant with respect to $\vec{w}$. This gives us a quadratic optimization problem that has a single global minimum. The solution to OLS is the unconstrained problem

$$\vec{w}^* = \underset{\vec{w}\in\mathbb{R}^n}{\operatorname{argmin}} f(\vec{w}) \tag{26.6}$$

Equation 26.6 can be solved by taking transposes and equating the stationary point of Equation 26.5 to the zero vector. Doing this gives us the normal equation

$$X^T X\vec{w}^* = X^T\vec{y} \tag{26.7}$$

Because X is assumed to be a full-rank matrix, $X^T X \succ 0$. The explicit solution to Equation 26.6, as a result of Equation 26.7, is

$$\vec{w}^* = [X^T X]^{-1} X^T\vec{y} \tag{26.8}$$

Practical Difficulties in OLS

In a later class, we will explore some of the practical shortcomings of the OLS solution. One foundational paper used the word "nonsensical" to describe the OLS solutions to real data that came from an industrial process.

Our primary concern is that OLS has a sensitivity to statistical outliers, which have quadratically increasing influence as they deviate from the optimal model. A secondary concern is that the error computed in an OLS estimate on training data are often not representative of the error that is subsequently computed on testing data. These concerns can be partially addressed by placing constraints on the solution vector $\vec{w}^*$ of OLS.

26.2 Constrained Least Squares

In constrained least squares (CLS), we place a single inequality constraint on the solution vector. One common constraint is to limit the magnitude of the solution. This requires the user to provide a *threshold value*, which we will write as the symbol θ.

The threshold value θ lets us impose a constraint on the magnitude of $\vec{w}$, such as $\|\vec{w}\|^2 \leq \theta$.

Our CLS problem is

$$
\begin{aligned}
\vec{w}^* &= \underset{\vec{w}\in\mathbb{R}^n}{\operatorname{argmin}}[X\vec{w}-\vec{y}]^T[X\vec{w}-\vec{y}] \\
&= \underset{\vec{w}\in\mathbb{R}^n}{\operatorname{argmin}}\,\vec{w}^T[X^TX]\vec{w}+[-2\vec{y}^TX]\vec{w} \qquad (26.9)\\
\text{such that} \quad \|\vec{w}\|^2 &\leq \theta
\end{aligned}
$$

The inequality in Equation 26.9 is a quadratic constraint, so it is a convex constraint. CLS, described by Equation 26.9, is therefore a convex problem: it has a convex objective and a single convex inequality constraint.

We can begin our solution by forming the Lagrange function

$$\mathcal{L}(\vec{w},\lambda) = f(\vec{w}) + \lambda(\|\vec{w}\|^2 - \theta) \tag{26.10}$$

The KKT conditions require that, for $\vec{w}^*$ to be a solution to Equation 26.9, the Lagrange multiplier is non-negative; mathematically, this is the requirement

$$\lambda^* \geq 0$$

The three other KKT conditions are primal feasibility, stationarity, and complementary slackness:

$$\|\vec{w}^*\|^2 \leq \theta \tag{26.11}$$

$$[\underline{\nabla}_{\vec{w}}\mathcal{L}(\vec{w}^*,\lambda^*)]^T = 2X^T[X\vec{w}^*-\vec{y}] + 2\lambda^*\vec{w}^* = 0 \tag{26.12}$$

$$\lambda(\|\vec{w}^*\|^2 - \theta) = 0 \tag{26.13}$$

The condition on complementary slackness is Equation 26.13, has two cases: the OLS solution is feasible or the OLS solution is not feasible. We can introduce a temporary abbreviation for the OLS solution as

$$\vec{w}^*_{LS} = [X^TX]X^T\vec{y}$$

The two cases of complementary slackness, and thus of primal feasibility, are detailed in the extra notes for this class. Our conclusion is that the OLS solution is feasible if and only if $\lambda = 0$.

Comments on CLS

Constrained least squares can be summarized as having this structure:

- If the OLS solution is feasible, then use the OLS solution
- Otherwise, first estimate the Lagrange multiplier λ^* and then estimate the optimal weight vector $\vec{w}^*$

Example: 9 data with 2 outliers

We can test the effect of CLS on a simple data set. Suppose that, for the independent data, we use as x_i the integers from 1 to 9. For dependent data we will use a formula based on Euler's number e:

$$\begin{aligned} y_1 &= e\,x_1 + \pi - 5 \\ y_i &= e\,x_i + \pi \qquad\qquad \text{for } i = 2 \dots 8 \\ y_9 &= e\,x_9 + \pi + 3 \end{aligned} \tag{26.14}$$

Our model of these data will be the first-order polynomial $X\vec{w} = \vec{c}$, in which X is the Vandermonde matrix, so the model is

$$\begin{bmatrix} \vec{x} & \vec{1} \end{bmatrix} \vec{w} = \vec{c}$$

If we fit a first-order polynomial model to all of the data, using ordinary least squares, we would get the weight vector

$$\vec{w}^*_{LS} \approx \begin{bmatrix} 3.1546 \\ 0.9780 \end{bmatrix}$$

If we use constrained least squares with $\theta = 8$ for a first-order polynomial model to all of the data, we would get the weight vector

$$\vec{w}^*_{CLS} \approx \begin{bmatrix} 2.7887 \\ 0.4724 \end{bmatrix}$$

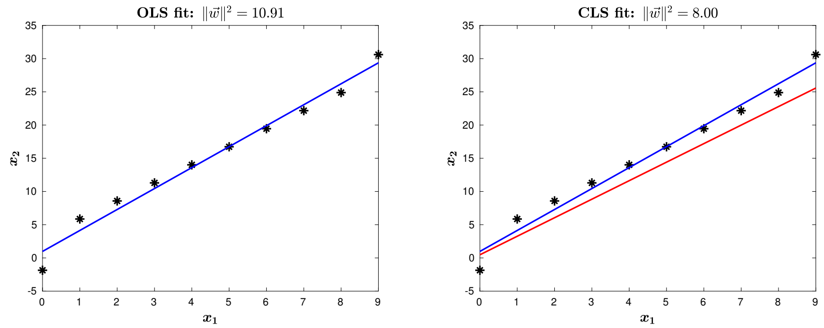
The RMS errors of these are

$$RMS(\vec{w}^*_{LS}) = 1.3376 \qquad\qquad RMS(\vec{w}^*_{CLS}) = 2.7434 \tag{26.15}$$

The results of Equation 26.15 are to be expected, because the OLS is the optimal fit to all of the data and the CLS – which includes a constraint on the norm of the weight vector – must have a greater overall error. The data and fits are graphically illustrated in Figure 26.1.

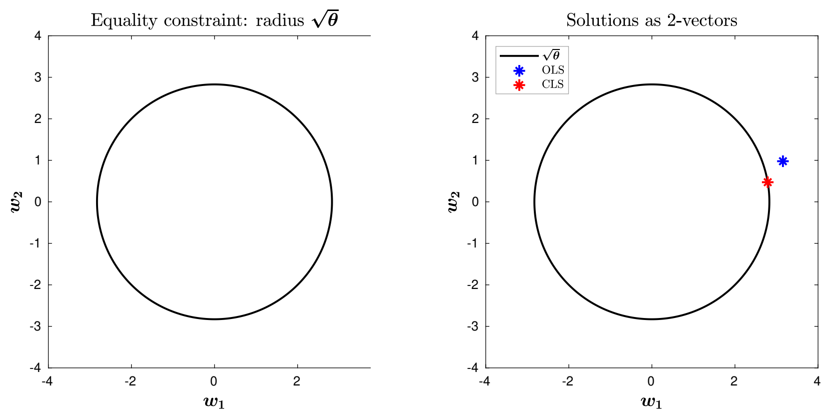
The effect of CLS, if needed, is to move the unconstrained OLS solution to satisfy the constraint on the weight vector. The weight vectors for the above example are have two entries, so the weight vectors are plotted in 2D in Figure 26.2. The OLS vector, shown as a red asterisk, is altered to the CLS solution that is indicated by the blue asterisk. We can see that the CLS solution does not simply scale the OLS solution so that the constraint is satisfied.

In the next class, we will briefly explore a motivation for using a constraint that reduces to the problem of CLS, but for which the value λ^* is *provided* rather than being *estimated*.



A **B**

Figure 26.1: Example data and least-squares fits of a first-order polynomial model to the data. (A) The data, as black asterisks, and the ordinary least-squares regression model as a red line. (B) The constrained least-squares solution using $\theta = 8$, shown as a blue line.



(A) **(B)**

Figure 26.2: Effect of imposing a constraint on a least-squares solution, for 2D data. (A) The equality constraint describes a circle in the vector space. (B)The ordinary least-squares solution produces a weight vector that is show as a red asterisk and the constrained least-squares solution produces a weight vector that is show as a blue asterisk.

26.3 Cross-Validation of Linear Regression

Let us assess these solutions by performing 10 passes of 5-fold cross-validation. We will combine RMS errors by taking the RMS of the RMS errors, which is the square root of the mean variance. When we do this for OLS, we get the data in Table 26.1.

Next, we will perform the same process using CLS with $\theta = 8$. These data are in Table 26.2.

Table 26.1: RMS of fits for the training subset and the testing subset for 5-fold cross-validation of ordinary least squares. The data are from Equation 26.14.

	Train	**Test**
	1.2713	1.8791
	1.2707	1.8738
	1.2629	1.9321
	1.2664	1.9071
	1.2517	2.0417
	1.2742	1.8745
	1.2668	1.9023
	1.2801	1.8134
	1.3084	1.6437
	1.2252	2.2373
Mean:	1.2678	1.9105
Std:	0.0210	0.1523

Table 26.2: RMS of training and testing for 5-fold cross-validation of constrained least squares, using $\theta = 8$. The data are from Equation 26.14.

	Train	**Test**
	2.7394	2.7834
	2.7411	2.7643
	2.7422	2.7546
	2.7418	2.7582
	2.7398	2.7781
	2.7429	2.7481
	2.7421	2.7546
	2.7400	2.7786
	2.7412	2.7666
	2.7388	2.7884
Mean:	2.7409	2.7675
Std.:	0.0014	0.0139

Observations on Table 26.1:

- For OLS, the test errors appear to be substantially greater that the training errors
  - The mean of OLS training fits is low and the variance is high, suggesting a *poor model* of the data
  - The mean of OLS tests is higher than the mean of training, and the variance is also higher, suggesting a *poor model* and *high variance* in tests

- For CLS, the training errors and the test errors appear to be comparable
  - The mean of CLS training errors is higher than the mean of OLS training errors and the variance of CLS training is an order of magnitude less that the OLS training variance, suggesting a *better model* using CLS
  - The mean of CLS tests is comparable to the mean of CLS training, and the variance of CLS tests is an order of magnitude lower than variance of CLS training, suggesting a *better model* and *low variance*

This example suggests that constrained least squares acts to reduce the variance of the tests, at the cost of a lower RMS training error to the given data.

Statistical explanations of an estimator's bias and variance have a long history, so many thorough analyses are available. The interested student is encouraged to explore the relationship between bias, variance, and intrinsic error for a model that includes a random variate with a Gaussian distribution.

26.4 Tikhonov Regularization

Andrey Tikhonov, in a seminal 1963 paper that is well explained in Joel Franklin's 1974 paper [3], was trying to solve difficult differential and integral equations. Tikhonov observed that, for a constraint value $\lambda > 0$ that is imposed on $\|\vec{w}\|^2$, the resulting equation is

$$g(\vec{w}, \lambda) = [X\vec{w} - \vec{y}]^T [X\vec{w} - \vec{y}] + \lambda \|\vec{w}\|^2 \tag{26.16}$$

We can see a close parallel between Equation 26.16 and the Lagrange function of Class #8, where Tikhonov's formulation has an crucial difference:

> *The Lagrange multiplier μ, which we computed in the optimization, is replaced by an argument λ that is <u>supplied</u> to us*

This difference is crucial because it created an entirely new way of viewing optimization problems, which Tikhonov called *regularization*.

We will write a slightly more general form so that we can solve some related problems.

Definition: Tikhonov regularization

For any full-rank matrix $X \in \mathbb{R}^{m \times n}$ with $m > n$, any vector $\vec{y} \in \mathbb{R}^n$, any full-rank matrix $R \in \mathbb{R}^{l \times n}$ with $l \leq n$, and any positive scalar $\alpha \in \mathbb{R}_{++}$, the *Tikhonov regularization function* is defined as

$$\mathcal{T}(\vec{w}, \alpha) \stackrel{\text{def}}{=} \|X\vec{w} - \vec{y}\|^2 + \|\alpha R\vec{w}\|^2 \tag{26.17}$$

We can see that Tikhonov's original function, in Equation 26.16, is Equation 26.17 using $R = I$ and $\lambda = \alpha^2$.

We can also see that the Lagrange equation for constrained least squares, in Equation 26.10, is very close in form to Tikhonov regularization. The differences in the equations are the arguments θ and α that a user must supply.

26.5 Regularization by Total Variation

A powerful method for analysis of functions is to study how much a function *varies* over an interval [1]. One interpretation of the total variation is that it is a measure of the absolute value of the derivative of a function. For example, under the usual assumptions in calculus, we know that the integral is the anti-derivative, so we have

$$w(t) = \int w'(t)dt$$

where the derivative of $w(t)$ is defined as

$$w'(t) = \lim_{h \to 0} \frac{w(t+h) - w(t)}{h}$$

If we convert the continuous function w into a discrete form, then we have a vector $\vec{w}$ that is of finite size n. The value $w(t)$ would be converted to w_i for a natural number i in the appropriate interval.

[1] Variation of an integral dates at least to Camille Jordan in 1881; the modern statement dates to 1937 by Stanislaw Saks [4], page 10.

The corresponding derivative would replace the continuous value $w(t+h)$ with a discrete w_{i+h}. To avoid division by zero, we would need to define the discrete derivative as

$$w_i' = \lim_{h \to 1} \frac{w_{i+h} - w_i}{h} = w_{i+1} - w_i \tag{26.18}$$

Under the usual assumptions about continuity and integrability, the *total variation* is the integral of the absolute value of the derivative. This is defined as

$$\int_{t_0}^{t_1} |w'(t)| dt \tag{26.19}$$

To change Equation 26.19 to a discrete form, we would replace the derivative term with the difference term in Equation 26.18, and replace the integral with a summation. Taking care to index the vector $\vec{w}$ with natural numbers in the interval $[1, n]$, the discrete version of total variation is

$$\sum_{i=1}^{n-1} |w_{i+1} - w_i| \tag{26.20}$$

In 1992, Rudin *et al.* [5] introduced the idea of using Equation 26.20 to remove noise from a 2D image. We will develop this for a 1D vector, from which higher dimensions can easily be inferred.

Since the publication of Rudin's work, researchers in image processing have determined that the absolute-value operator in Equation 26.20 can be replaced with a squaring operator. This does not substantially change the effectiveness of most results and it produces a much simpler equation. We will measure the *square* of the total variation of the solution vector $\vec{w}$. Because $(w_{i+1} - w_i)^2 = (w_i - w_{i+1})^2$, we can modify Equation 26.20 and change the indexing to use the measure

$$\sum_{i=1}^{n-1} (w_i - w_{i+1})^2 \tag{26.21}$$

A useful simplification of Equation 26.21 is to observe th repeating circulant pattern of differences. We can write the vector of differences using a non-square matrix R that is, depending on our point of view, a Toeplitz matrix or a circulant matrix. The differences are

$$\begin{bmatrix} w_1 - w_2 \\ w_2 - w_3 \\ w_3 - w_4 \\ \vdots \\ w_{n-1} - w_n \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} \tag{26.22}$$

$$\equiv \qquad \vec{w}\,' = R\vec{w}$$

The squared variation of Equation 26.21 can be written, in matrix form, as

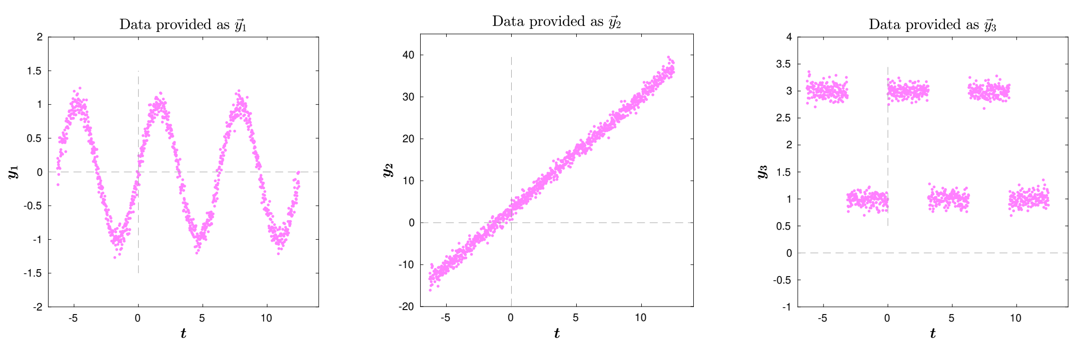
$$\sum_{i=1}^{n-1} (w_i - w_{i+1})^2 = [R\vec{w}]^T [R\vec{w}] = \|R\vec{w}\|^2 \tag{26.23}$$

26.6 Tikhonov Regularization for Denoising

Suppose that we want to estimate a vector $\vec{w}$ that has been altered by the addition of another vector. The "other vector" is often called "noise", and this assumption is called *additive noise*. We assume that a given data vector $\vec{y}$ is the unknown vector $\vec{w}$ plus an unknown noise vector $\vec{v}$, so

$$\vec{y} = \vec{w} + \vec{v} \tag{26.24}$$

Examples of this kind of alteration to an ideal vector are shown in Figure 26.3.



(A) **(B)** **(C)**

Figure 26.3: Three signals to which a normally distributed variate has been added. (A) Underlying signal is a sinusoid. (B) Underlying signal is linear. (C) Underlying signal is a square wave.

A reasonable objective in solving Equation 26.24 for $\vec{w}$, given a data vector $\vec{y}$, is to minimize the squared residual vector. Our objective function, disregarding terms that do not mention $\vec{w}$, is

$$f(\vec{w}) = \|\vec{w} - \vec{y}\|^2 = [\vec{w} - \vec{y}]^T[\vec{w} - \vec{y}] = \vec{w}^T\vec{w} - 2\vec{y}^T\vec{w} \tag{26.25}$$

Equation 26.23 is in the form of the regularization term in Tikhonov regularization, so we can form a function according to Equation 26.17. The matrix is $X = I$. We can set $\lambda = \alpha^2$ and take the Tikhonov argument out of the vector norm. This give us the expanded form of the Tikhonov function as

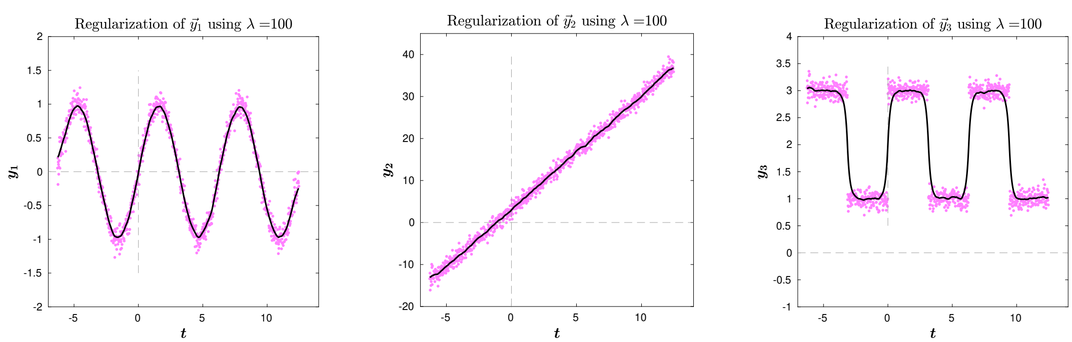
$$\mathcal{T}(\vec{w}, \lambda) = \vec{w}^T\vec{w} - 2\vec{y}^T\vec{w} + \lambda\vec{w}^T R^T R\vec{w} \tag{26.26}$$

We can take the derivative of Equation 26.26 with respect to the vector $\vec{w}$ and, to find the stationary point, set the transpose equal to the zero vector. This gives the linear equation

$$\begin{aligned} 2\vec{w}^* - 2\vec{y} + 2\lambda R^T R\vec{w}^* &= \vec{0} \\ \equiv \quad \left[I + \lambda R^T R\right]\vec{w}^* &= \vec{y} \end{aligned} \tag{26.27}$$

Equation 26.27 can be solved using ordinary methods from linear algebra.

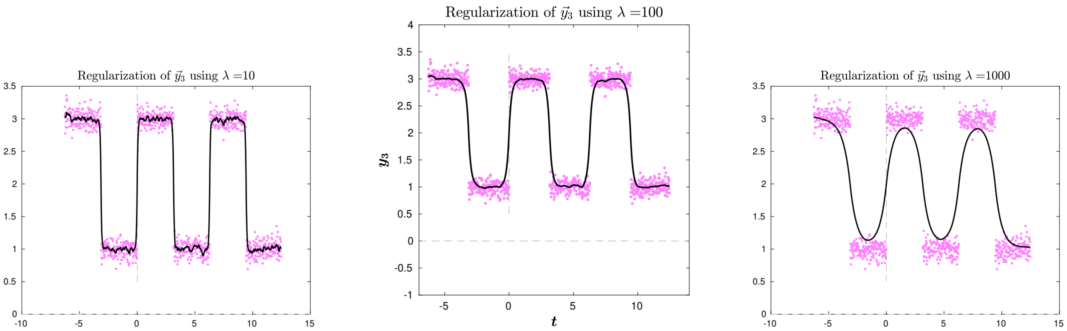
We can use Equation 26.27, with $\lambda = 100$, to regularize the data presented in Figure 26.3. The results of Tikhonov regularization are presented in Figure 26.4.



(A) **(B)** **(C)**

Figure 26.4: Tikhonov regularization for denoising three signals $\vec{w}$, using argument value $\lambda = 100$. (A) Underlying $\vec{w}$ is a sinusoid. (B) Underlying $\vec{w}$ is linear. (C) Underlying $\vec{w}$ is a square wave.

The value of the regularization argument λ must be positive, and its numerical value makes a substantial difference to the results. The additive noise in the data was normally distributed, with an amplitude of 0.1; the ideal value for λ is the inverse of the square of the noise amplitude, and hence $\lambda = 100$ was selected. Results for other values of λ are shown in Figure 26.5.



(A) **(B)** **(C)**

Figure 26.5: Tikhonov regularization for denoising a square-wave signal $\vec{w}$, using three argument values for λ. (A) With $\lambda = 10$, the fit contains high-frequency components. (B) With $\lambda = 100$, the fit is close to the underlying function. (C) With $\lambda = 100$, the fit has a high error to the data.

Figure 26.5 is a visual presentation of a common observation: the argument λ is a *smoothing* hyper-parameter, and the results vary from less smooth to more smooth as λ increases.

Extra Notes

26.7 Extra Notes on Constrained Least Squares

As described in the main notes for this class, in CLS either the OLS solution is feasible in the primal formulation or the OLS solution is infeasible. We can reason about these two cases to determine what action we need to take.

Case 1: $\vec{w}^*_{LS}$ is feasible

If the OLS solution is feasible, then the constraint is inactive and we can set the Lagrange multiplier $\lambda = 0$. This is dual feasible and satisfies complementary slackness. Thus, if the OLS solution is feasible, then the Lagrange multiplier $\lambda = 0$. We can also reason that the converse is true.

If $\lambda = 0$, then the complementary slackness constraint of Equation 26.13 reduces to the Lagrange equation for the OLS problem. The solution to the CLS problem would then be the solution to the OLS problem, which is $\vec{w}^*$. If $\vec{w}^*$ satisfies Equation 26.11, then $\vec{w}^*$ is feasible.

Case 2: $\vec{w}^*_{LS}$ is not feasible

We have reasoned, above, that the OLS solution is feasible if and only if $\lambda^* = 0$. The second case is when the OLS solution is not feasible, which must – by our reasoning – be equivalent to $\lambda^* > 0$. We can write Equation 26.12, the constraint on stationarity, in a way that lets us find a solution for $\vec{w}^*$ in terms of λ^*. We can do this as

$$\begin{array}{rrcl} & 2X^T[X\vec{w}^* - \vec{y}] + 2\lambda^*\vec{w}^* & = & 0 \\ \equiv & X^TX\vec{w}^* - X^T\vec{y} + \lambda^* I\vec{w}^* & = & 0 \\ \equiv & [X^TX + \lambda^* I]\vec{w}^* & = & X^T\vec{y} \\ \equiv & \vec{w}^* & = & \left[X^TX + \lambda^* I\right]^{-1} X^T\vec{y} \\ \equiv & \vec{w}^*(\lambda^*) & = & \left[X^TX + \lambda^* I\right]^{-1} X^T\vec{y} \end{array} \tag{26.28}$$

According to Equation 26.28, the optimal weight vector $\vec{w}^*$ is a function of the optimal Lagrange multiplier λ^*. We can compute λ^* by requiring that both the optimal Lagrange multiplier and the optimal weight vector must satisfy the dual feasibility constraint, which is Equation 26.13. Because we are in the case where $\lambda > 0$, the constraint is satisfied if and only if $\|\vec{w}^*(\lambda^*)\|^2 - \theta = 0$.

We can write a new, temporary, function

$$g(\lambda^*) = \|\vec{w}^*(\lambda^*)\|^2 - \theta = 0 \tag{26.29}$$

We can solve Equation 26.29 numerically using many methods, including the builtin `MATLAB` function `fzero` using the hyper-parameter $\lambda^* = 0$ as an initial estimate. Another method for estimating λ^* is to bracket the root of $g(\lambda^*)$ and use a bisection search for the root of the function.

Using a zero-finding method, a bracketing method, or another appropriate method, from λ^* we can compute $\vec{w}^* = \vec{w}^*(\lambda^*)$. This completes the second case of CLS.

End of Extra Notes

References

[1] Beck A: Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB. Siam Press, 2014

[2] Boyd S, Vandenberghe L: Convex Optimization. Cambridge University Press, 2004

[3] Franklin JN: On Tikhonov's method for ill-posed problems. *Math Comput* 28(128):889–907, 1974

[4] Saks S: Theory of the Integral, 2d rev. ed. Dover Publications, New York, 1964

[5] Rudin LI, Osher S, Fatemi E: Nonlinear total variation based noise removal algorithms. *Physica D* 60(1-4):259–268, 1992