51 REGULARIZATION

We discussed the least-squares problem in the last chapter, which uses a collection of data points $\{x(n), y_n\}$ to determine an optimal parameter $w^\star$ by minimizing an empirical quadratic risk of the form:

$$w^\star = \underset{w\in\mathbb{R}^M}{\operatorname{argmin}} \left\{ P(w) \triangleq \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n) - y_n^{\mathsf{T}} w\Big)^2 \right\} \tag{51.1a}$$

where each y_n is $M-$dimensional and each $x(n)$ is a scalar. The solution is determined by solving the normal equations:

$$H^{\mathsf{T}} H w^\star = H^{\mathsf{T}} d, \qquad (\textbf{normal equations}) \tag{51.1b}$$

where the quantities $d \in \mathbb{R}^{N\times 1}$ and $H \in \mathbb{R}^{N\times M}$ collect the data:

$$H \triangleq \begin{bmatrix} y_0^{\mathsf{T}} \\ y_1^{\mathsf{T}} \\ y_2^{\mathsf{T}} \\ \vdots \\ y_{N-1}^{\mathsf{T}} \end{bmatrix}, \qquad d \triangleq \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(N-1) \end{bmatrix} \tag{51.1c}$$

The normal equations (51.1b) may have a unique solution or infinitely many solutions. They may also be ill-conditioned meaning that slight perturbations to the data $\{d, H\}$ can lead to large changes in the solution $w^\star$; this usually occurs when the matrix H is ill-conditioned. In this chapter, we will use the least-squares formulation as a guiding example to illustrate three types of challenges that arise in data-driven learning methods pertaining to **(a)** non-uniqueness of solutions, **(b)** ill-conditioning, and **(c)** the undesirable possibility of over-fitting. We will then explain that regularization is a useful tool to alleviate these challenges. We will also explain how regularization enables the designer to promote preference for certain solutions such as favoring solutions with small norms or sparse structure. We will motivate the main ideas by using the least-squares formulation due to its mathematical tractability. Subsequently, we will extend the discussion more general empirical risks, other than the least-squares case, which will arise in later chapters when we deal with logistic regression, support vector machines, kernel machines, neural networks, and other learning methods.

51.1 THREE CHALLENGES

In learning problems, we make a distinction between *training* data and *test* data. The data $\{x(n), y_n\}$ used to solve the least-squares problem (51.1a) are referred to as *training data*. Once a solution $w^\star$ is determined, the value of the risk function at the solution is called the *training error*:

$$\textbf{training error} \triangleq \frac{1}{N}\sum_{n=0}^{N-1}(x(n) - y_n^{\mathsf{T}} w^\star)^2 \;=\; \frac{1}{N}\sum_{n=0}^{N-1}(x(n) - \widehat{x}(n))^2 \tag{51.2}$$

where $\widehat{x}(n) = y_n^{\mathsf{T}} w^\star$ denotes the prediction for $x(n)$. In this way, the training error is measuring how well the least-squares solution performs on the training data. In general, the training error will be small because the solution $w^\star$ is purposefully determined to minimize it.

In most learning applications, however, the main purpose for learning $w^\star$ is to employ it to perform prediction on future data that *were not* part of the training phase. For this reason, it is customary to assess performance on a separate collection of T test data points denoted by $\{x(t), y_t\}$, and which are assumed to arise from the same underlying distribution $f_{\boldsymbol{x},\boldsymbol{y}}(x, y)$ as the training data. The corresponding testing error is defined by

$$\textbf{testing error} \triangleq \frac{1}{T}\sum_{t=0}^{T-1}(x(t) - y_t^{\mathsf{T}} w^\star)^2 \;=\; \frac{1}{T}\sum_{t=0}^{T-1}(x(t) - \widehat{x}(t))^2 \tag{51.3}$$

where $\widehat{x}(t) = y_t^{\mathsf{T}} w^\star$ denotes the prediction for $x(t)$. In general, the testing error will be larger than the training error but we desire the gap between them to be small. Learning algorithms that lead to small error gaps are said to *generalize* well, namely, they are able to extend their good performance on training data to the test data as well. We will discuss generalization and training and testing errors in greater detail in future chapters, especially in the context of classification problems. Here, we are using the least-squares problem to motivate the concepts.

Difficulties

We already know that the normal equations (51.1b) are consistent, meaning that a solution $w^\star$ always exists. The solution is either unique when H has full column rank, in which case it is given by

$$w^\star = (H^{\mathsf{T}} H)^{-1} H^{\mathsf{T}} d, \quad (H \text{ has full column rank}) \tag{51.4}$$

or there are infinitely many solutions differing by vectors in $\mathcal{N}(H)$. Some challenges arise in both scenarios, which lead to complications when solving inference problems:

(a) (**Non-uniqueness**). When infinitely many solutions exist, the training error

will not change regardless of which solution we pick. This is because all valid solutions $w^\star$ differ by vectors in the null space of H and, therefore, if $w_1^\star$ and $w_2^\star$ are two valid solutions then

$$w_2^\star = w_1^\star + p, \quad \text{for some } p \in \mathcal{N}(H) \tag{51.5}$$

In this case, the predictions $\widehat{x}(n)$ for the training signals will remain unchanged under $w_1^\star$ or $w_2^\star$ since $Hp = 0$ and, hence, $y_n^{\mathsf{T}} p = 0$ for any of the observation vectors in the training set so that

$$\widehat{x}(n) = y_n^{\mathsf{T}} w_2^\star = y_n^{\mathsf{T}} w_1^\star \tag{51.6}$$

It follows that the training error remains invariant. However, the testing error will be sensitive to which solution we select because the *test* observations $\{y_t\}$ need not be orthogonal anymore to the nullspace of H. We explain in the sequel that ℓ_2-regularization forces a unique solution $w^\star$ and removes this ambiguity.

(b) (**Overfitting**) Infinitely many solutions $w^\star$ can exist even when $N \geq M$, i.e., even when we have more observations than unknown entries. This occurs when the columns of H are linearly dependent and gives rise a second challenge. Recall that the least-squares problem is approximating d by $\widehat{d} = Hw^\star$. When the columns of H are linearly dependent, some of its columns can be removed to obtain a full-rank lower-dimensional matrix, $H' \in \mathbb{R}^{N \times M'}$ with $M' < M$. This new matrix spans the same column space as H:

$$\mathcal{R}(H') = \mathcal{R}(H) \tag{51.7}$$

We can then solve an equivalent least-squares problem involving $\{d, H'\}$ instead of $\{d, H\}$ to obtain the same projection $\widehat{d}$ by using a smaller-size solution $(w')^\star$ of dimension M'. We thus see that the rank-deficiency of H amounts to using a more complex model w (i.e., of higher dimensions) than is necessary to approximate d. This issue is a manifestation of the problem of *overfitting*, which we will discuss in greater detail in later chapters. Overfitting amounts to using more complex models than necessary and it also degrades performance on test data.

Rank-deficiency of H also arises when $N < M$ (i.e., when H has more columns than rows). One way to deal with this problem is to collect more data (i.e., to use a larger N). A second way is to perform dimensionality reduction and reduce the size of the observation vectors. We will discuss techniques for dimensionality reduction in later chapters, including the principal component analysis (PCA) method and the Fisher discriminant analysis (FDA) method. A third way is to employ regularization. For example, we will explain further ahead that ℓ_1-regularization automatically selects a subset of the columns of H to compute $w^\star$.

(c) (**Ill-conditioning**) Difficulties can arise even when the normal equations have a unique solution $w^\star$ but the data matrix H is ill-conditioned (i.e., has a large

condition number). In this case, small changes in the data $\{d, H\}$ can lead to large changes in the solution $w^\star$ and affect the inference conclusion and testing error — see Prob. 51.2 for a numerical example. One leading cause for ill-conditioning is when the entries within the observation vectors are not normalized properly so that some entries are disproportionately larger by some orders of magnitude than other entries. Such large discrepancies can distort the operation of a learning algorithm, including the least-squares solution, by giving more relevance or attention to larger entries in the observation vector over other entries. One way to deal with ill-conditioning is therefore to *scale* the observation vectors so that their entries assume values within some uniform range. The next example explains how scaling can be performed. A second way is to employ regularization. In particular, we will see that ℓ_2-regularization reduces the effect of ill-conditioning.

Example 51.1 (**Normalization of observation vectors**) It is common practice to center the training data around their sample means, as was already suggested by the discussion in Sec. 29.2. We can take this step further and normalize the entries of the observation vectors to have unit-variance as well. Specifically, the first step is to compute the sample mean vector:

$$\bar{y} \triangleq \frac{1}{N}\sum_{n=0}^{N-1} y_n \tag{51.8a}$$

and to use it to center all observation vectors by replacing them by

$$y_{n,c} \triangleq y_n - \bar{y} \tag{51.8b}$$

where, for clarity, we are adding the subscript "c" to refer to centered variables. If we denote the individual entries of $\{\bar{y}, y_n\}$ by $\{\bar{y}(m), y(m),\ m = 1, 2, \ldots, M\}$, then centering amounts to replacing the individual entries by

$$y_{n,c}(m) \triangleq y_n(m) - \bar{y}(m) \tag{51.8c}$$

The second step in the normalization process is to evaluate the (unbiased) sample variance for each of these centered entries, namely,

$$\widehat{\sigma}_m^2 \triangleq \frac{1}{N-1}\sum_{n=0}^{N-1} y_{n,c}^2(m),\ \ m = 1, 2, \ldots, M \tag{51.9a}$$

and to scale $y_{n,c}(m)$ by the corresponding standard deviation to get

$$y_{n,p}(m) \triangleq y_{n,c}(m)/\widehat{\sigma}_m,\ \ m = 1, 2, \ldots, M \tag{51.9b}$$

where we are now using the subscript "p." In this way, we start from an observation vector y_n and replace it by the normalized vector $y_{n,p}$, where all entries of $y_{n,p}$ are centered with zero mean and unit variance:

$$\{y_n\} \xrightarrow{\text{remove sample mean}} \{y_{n,c}\} \xrightarrow{\text{normalize variance}} \{y_{n,p}\} \tag{51.10}$$

A second method to normalize the observation vectors $\{y_n\}$ is as follows. We first identify the smallest and largest entry values within the given dataset:

$$y_{\min} \triangleq \min_{n,m} y_n(m) \tag{51.11a}$$

$$y_{\max} \triangleq \max_{n,m} y_n(m) \tag{51.11b}$$

$$\Delta = y_{\max} - y_{\min} \tag{51.11c}$$

and then scale all entries in the following manner, for each n and m:

$$y_{n,s}(m) \triangleq \frac{y_n(m) - y_{\min}}{\Delta} \tag{51.12}$$

In this way, each scaled entry $y_{n,s}(m)$ will assume values within the range $[0, 1]$. We can subsequently center the means of these entries at zero by computing

$$y_{n,p} \triangleq y_{n,s} - \bar{y}_{n,s}, \quad \text{where} \quad \bar{y}_{n,s} = \frac{1}{N}\sum_{n=0}^{N-1} y_{n,s} \tag{51.13}$$

Here again, we start from a given observation vector y_n and replace it by $y_{n,p}$, where all entries lie within the range $[-1, 1]$:

$$\{y_n\} \xrightarrow{\text{normalize range}} \{y_{n,s}\} \xrightarrow{\text{remove sample mean}} \{y_{n,p}\} \tag{51.14}$$

Regardless of which normalization procedure is used, we will assume that the given observation vectors $\{y_n\}$ have already gone through this process and will continue to use the notation y_n rather than switch to $y_{n,p}$ for simplicity.

51.2 ℓ_2-REGULARIZATION

One useful technique to avoid the challenges of non-uniqueness of solutions, overfitting, and ill-conditioning is to employ *regularization* (also called *shrinkage* in the statistics literature). The technique penalizes some norm of the parameter w in order to favor solutions with desirable properties based on some prior knowledge (such as sparse solutions or solutions with small Euclidean norm). We say that regularization incorporates a form of *inductive bias* in that it biases the solution away from the unregularized case by incorporating some prior information. This is attained by adding an explicit convex penalty term to the original risk function such as

$$q(w) = \begin{cases} \rho\|w\|^2, & (\ell_2-\text{regularization}) \\ \alpha\|w\|_1, & (\ell_1-\text{regularization}) \\ \alpha\|w\|_1 + \rho\|w\|^2, & (\text{elastic-net regularization}) \\ \beta\|w\|_0, & (\ell_0-\text{ regularization}) \end{cases} \tag{51.15}$$

where (α, β, ρ) are nonnegative parameters, and where $\|w\|_0$ is a pseudo-norm that counts the number of nonzero elements in w. We will focus on the first

three choices due to their mathematical tractability. One can also consider other vector norms, such as the $p-$th norm, $\|w\|_p$ for $p < 1$ or $p = \infty$. Regularization will generally have a limited effect on the *training error* of an algorithm, but will improve the *generalization* ability of the algorithm by improving its performance on test data for the reasons explained in the sequel. We consider first the case of ℓ_2-regularization, also called *ridge regression*, where the penalty term is quadratic in w.

51.2.1 Ridge Regression

In ridge regression, we replace the empirical risk (51.1a) by the regularized version:

$$w^\star_{\rm reg} \triangleq \operatorname*{argmin}_{w\in\mathbb{R}^M} \left\{ P_{\rm reg}(w) \triangleq \rho\|w\|^2 + \frac{1}{N}\sum_{n=0}^{N-1} \big(x(n) - y_n^{\mathsf{T}} w\big)^2 \right\} \tag{51.16}$$

where $\rho > 0$ is the regularization factor; its value may or may not depend on N. In general, the value of ρ is independent of N.

Observe that, for the purposes of this chapter, we are adding a subscript "reg" to $(w^\star_{\rm reg}, P_{\rm reg}(w))$ to distinguish them from the unregularized versions $(w^\star, P(w))$. This is because we will be comparing both risks and their minimizers throughout this chapter. In future chapters, however, where we will be working almost exclusively with regularized risks, we will revert to the traditional notation $(w^\star, P(w))$ without the "reg" subscript for simplicity. Before explaining how ridge regression addresses the aforementioned challenges, we revisit Example 50.1 and show how the regularized empirical risk (51.16) can be motivated as the solution to a maximum a-posteriori (MAP) inference problem.

Example 51.2 (**Interpretation in terms of a Gaussian prior on the model**) Assume we collect N independent and identically-distributed observations $\{\boldsymbol{x}(n),\ y_n\}$, for $0 \le n \le N-1$. Assume also that these observations satisfy the same linear model (50.20), namely,

$$\boldsymbol{x}(n) = y_n^{\mathsf{T}} w + \boldsymbol{v}(n) \tag{51.17}$$

for some unknown $w \in \mathbb{R}^M$, and where $\boldsymbol{v}(n)$ is a white Gaussian noise process with zero mean and variance σ_v^2. In the earlier Example 50.1, the model w was treated as an *unknown constant* and a maximum-likelihood formulation was used to estimate it; thus leading to the standard least-squares problem. Here, we will instead model w as a realization for some random variable $\boldsymbol{w}$ that is Gaussian-distributed with zero mean and covariance matrix $R_w = \sigma_w^2 I_M$, i.e.,

$$f_{\boldsymbol{w}}(w) = \frac{1}{\sqrt{(2\pi\sigma_w^2)^M}} \exp\Big\{-\frac{1}{2\sigma_w^2}\|w\|^2\Big\} \tag{51.18}$$

Once w is selected from this distribution, then all observations $\{\boldsymbol{x}(n)\}$ are generated by this *same* w from knowledge of $\{x(n), y_n\}$. We are again interested in estimating w. Using Bayes rule (3.39), we assess the conditional probability distribution of the model

given the observations as follows:

$$\begin{aligned}
&f_{\boldsymbol{w}|\boldsymbol{x},\boldsymbol{y}}\Big(w|\{x(n),y_n\}\Big)\\
&\propto f_{\boldsymbol{x},\boldsymbol{y}|\boldsymbol{w}}\Big(\{x(n),y_n\}\,|\,w\Big)\,f_{\boldsymbol{w}}(w)\\
&=\left\{\prod_{n=0}^{N-1} f_{\boldsymbol{v}}\Big(x(n)-y_n^{\mathsf{T}}w\Big)\right\}f_{\boldsymbol{w}}(w)\\
&\propto\left\{\prod_{n=0}^{N-1}\exp\Big\{-\frac{1}{2\sigma_v^2}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\Big\}\right\}\times\exp\left\{-\frac{1}{2\sigma_w^2}\|w\|^2\right\}\\
&=\exp\left\{-\frac{1}{2\sigma_w^2}\|w\|^2-\frac{1}{2\sigma_v^2}\sum_{n=0}^{N-1}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\}
\end{aligned}\tag{51.19}$$

where the first and third lines replace the equality sign by proportionality constants. Consequently, we can now formulate a *maximum a-posteriori* (MAP) estimation problem to recover w, which amounts to seeking the value of w that maximizes the above conditional density function:

$$\begin{aligned}
w_{\rm reg}^{\star} &\triangleq \operatorname*{argmax}_{w\in\mathbb{R}^M}\; f_{\boldsymbol{w}|\boldsymbol{x},\boldsymbol{y}}\Big(w|\{x(n),y_n\}\Big)\\
&=\operatorname*{argmin}_{w\in\mathbb{R}^M}\;\left\{\frac{1}{2\sigma_w^2}\|w\|^2+\frac{1}{2\sigma_v^2}\sum_{n=0}^{N-1}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\}\\
&=\operatorname*{argmin}_{w\in\mathbb{R}^M}\;\frac{N}{2\sigma_v^2}\left\{\frac{2\sigma_v^2}{2N\sigma_w^2}\|w\|^2+\frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\}\\
&=\operatorname*{argmin}_{w\in\mathbb{R}^M}\;\left\{\rho\|w\|^2+\frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\}
\end{aligned}\tag{51.20}$$

where we introduced $\rho=\sigma_v^2/N\sigma_w^2$. We therefore recover the regularized empirical risk (51.16). This argument shows that ℓ_2-regularization helps ensure that the solution $w_{\rm reg}^{\star}$ is consistent with a prior Gaussian model on the distribution of $\boldsymbol{w}$.

We now explain how ridge regression promotes solutions with smaller Euclidean norm and alleviates the challenges of ill-conditioning, over-fitting, and non-uniqueness of solutions.

Resolving non-uniqueness

Differentiating $P_{\rm reg}(w)$ in (51.16) with respect to w, we find that the solution is unique and given by

$$w_{\rm reg}^{\star}\;=\;(\rho NI_M+H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}d\tag{51.21}$$

where the matrix $\rho NI_M+H^{\mathsf{T}}H$ is always invertible due to the positive term, $\rho NI_M>0$ and independent of whether H is rank-deficient or not.

Promoting smaller solutions

It is seen from the regularized risk (51.16) that larger values for ρ favor solutions $w_{\rm reg}^{\star}$ with smaller Euclidean norm than would result when $\rho=0$. This is because

the objective is to minimize the aggregate risk, and the first term is influenced by $\rho\|w\|^2$. This property can be established more formally as follows (see Prob. 51.3 for an alternative argument). Using the unregularized risk $P(w)$, and since $w^\star_{\rm reg}$ minimizes the regularized risk, we have

$$\begin{aligned}
&\rho\|w^\star_{\rm reg}\|^2 + P(w^\star_{\rm reg}) \le \rho\|w^\star\|^2 + P(w^\star) \\
&\Longrightarrow \rho\|w^\star_{\rm reg}\|^2 - \rho\|w^\star\|^2 \le P(w^\star) - P(w^\star_{\rm reg}) \\
&\overset{(a)}{\Longrightarrow} \rho\|w^\star_{\rm reg}\|^2 - \rho\|w^\star\|^2 \le 0
\end{aligned} \tag{51.22}$$

where step (a) is because $w^\star$ minimizes the unregularized risk, $P(w)$. It follows that $\|w^\star_{\rm reg}\|^2 \le \|w^\star\|^2$. Actually, strict inequality holds because $P(w^\star)$ is strictly smaller than $P(w^\star_{\rm reg})$. Since, otherwise, for $P(w)$ to assume the same value at both $(w^\star, w^\star_{\rm reg})$, it would mean that $w^\star_{\rm reg}$ must be a minimizer for $P(w)$ as well. In that case, both $(w^\star, w^\star_{\rm reg})$ must satisfy the *same* normal equations, namely,

$$H^{\sf T}Hw^\star = H^{\sf T}d, \quad H^{\sf T}Hw^\star_{\rm reg} = H^{\sf T}d \tag{51.23}$$

But since $w^\star_{\rm reg}$ satisfies (51.21), i.e., $()\rho NI_M + H^{\sf T}H)w^\star_{\rm reg} = H^{\sf T}d$, we conclude that $w^\star_{\rm reg} = 0$. But this is not possible unless $H^{\sf T}d = 0$. Absent this condition, we conclude that

$$\|w^\star_{\rm reg}\|^2 < \|w^\star\|^2 \tag{51.24}$$

This proves that the norm of the regularized solution, $w^\star_{\rm reg}$, shrinks in comparison to the norm of the original solution, $w^\star$. This property is referred to as *shrinkage*. We will encounter it in other regularization formulations as well.

Countering ill-conditioning

Regularization also counters the effect of ill-conditioning, i.e., the sensitivity of the solution $w^\star$ to small variations in the data $\{x(n), y_n\}$. Note that the condition number of the new coefficient matrix is given by

$$\kappa(\rho NI_M + H^{\sf T}H) \;\triangleq\; \frac{\rho N + \lambda_{\max}(H^{\sf T}H)}{\rho N + \lambda_{\min}(H^{\sf T}H)} \;=\; \frac{\rho N + \sigma^2_{\max}(H)}{\rho N + \sigma^2_{\min}(H)} \tag{51.25}$$

in terms of the largest and smallest singular values of H. If the value of ρN is large enough in comparison to the singular-value spread of H, then the ratio on the right-hand side approaches one and the matrix $\rho NI_M + H^{\sf T}H$ becomes (very) well conditioned.

Countering overfitting

By promoting solutions $w^\star_{\rm reg}$ with smaller Euclidean norm, regularization helps alleviate the danger of overfitting because it searches for the solution over a reduced region in space. This can be shown more formally by verifying that minimizing a regularized least-squares problem of the form (51.16) is equivalent to solving a *constrained* optimization problem of the following form:

$$w^\star_{\rm reg} \;\triangleq\; \underset{w\in\mathbb{R}^M}{\rm argmin} \left\{ \frac{1}{N}\sum_{n=0}^{N-1}(x(n) - y_n^{\sf T}w)^2 \right\}, \quad \text{subject to } \rho\|w\|^2 \le \tau \tag{51.26}$$

for some $\tau > 0$. The equivalence between problems (51.16) and (51.26) is established algebraically in Appendix 51.A by using the Lagrange and KKT multiplier arguments from Sec. 9.1. This equivalent characterization shows that regularization reduces the search space for w to the spherical region $\|w\|^2 \leq \tau/\rho$ instead of searching over the entire space $w \in \mathbb{R}^M$. Some care is needed in selecting ρ (or τ): large values for ρ (or small τ) can have the opposite effect and constrain the search region excessively, thus leading to the possibility of *underfitting* (i.e., to the use of simpler models than is actually necessary to fit the data well). These remarks show that there is a compromise in setting the value of ρ: small ρ does not perform effective regularization and large ρ can cause underfitting.

Biased risk values

Although regularization is effective in countering ill-conditioning and overfitting, there is a price to pay. This is because regularization biases the least attainable risk (i.e., the training error), which becomes larger than in the unregularized case. To see this, consider again the solutions $w^\star$ and $w^\star_{\rm reg}$ to the standard and regularized least-squares problems. Evaluating the risk functions at the respective minimizers and subtracting them we get, after some algebra — see Prob. 51.4:

$$P_{\rm reg}(w^\star_{\rm reg}) \;-\; P(w^\star) = \rho\,(w^\star)^{\mathsf{T}} w^\star_{\rm reg} \;> 0 \tag{51.27}$$

from which we conclude that $P_{\rm reg}(w^\star_{\rm reg}) > P(w^\star)$, and that the bias increases with ρ.

Example 51.3 (**QR solution method**) Determination of the ℓ_2-regularized solution (51.21) requires that we compute the matrix product $H^{\mathsf{T}}H$ and invert the matrix $\rho N I_M + H^{\mathsf{T}}H$. We explained earlier in Prob. 50.5 that squaring matrix entries through the product $H^{\mathsf{T}}H$ can lead to a loss in numerical precision for small entries; it can also lead to overflow for large entries. A more stable numerical procedure for determining $w^\star_{\rm reg}$ can be motivated by using the QR decomposition. We construct the extended quantities:

$$H^e \;\triangleq\; \begin{bmatrix} H \\ \sqrt{\rho N} I_M \end{bmatrix}, \quad \text{of size } (N+M)\times M \tag{51.28}$$

$$d^e \;\triangleq\; \begin{bmatrix} d \\ 0_{M\times 1} \end{bmatrix}, \quad \text{of size } (N+M)\times 1 \tag{51.29}$$

and introduce the QR decomposition:

$$H^e = Q \begin{bmatrix} R \\ 0 \end{bmatrix} \tag{51.30}$$

where

$$R:\ (M\times M), \quad Q:\ (N+M)\times(N+M), \quad QQ^{\mathsf{T}} = Q^{\mathsf{T}}Q = I \tag{51.31}$$

and R is upper triangular. We apply the orthogonal transformation Q^{T} to d^e and denote the resulting entries by

$$Q^{\mathsf{T}} d^e = \begin{bmatrix} \bar{d} \\ \times \end{bmatrix}, \quad \bar{d}:\ (N\times 1) \tag{51.32}$$

where $\times$ refers to irrelevant entries. Then, note from (51.28) that

$$(H^e)^{\mathsf{T}} H^e = \rho N I_M + H^{\mathsf{T}} H \tag{51.33}$$

while from (51.30)

$$(H^e)^{\mathsf{T}} H^e = \begin{bmatrix} R^{\mathsf{T}} & 0 \end{bmatrix} Q^{\mathsf{T}} Q \begin{bmatrix} R \\ 0 \end{bmatrix} = R^{\mathsf{T}} R \tag{51.34}$$

It then follows that

$$\begin{aligned}
w_{\rm reg}^{\star} &= (\rho N I_M + H^{\mathsf{T}} H)^{-1} H^{\mathsf{T}} d \\
&= \left((H^e)^{\mathsf{T}} H^e\right)^{-1} (H^e)^{\mathsf{T}} \begin{bmatrix} d \\ 0 \end{bmatrix} \qquad (51.35) \\
&= R^{-1} (R^{\mathsf{T}})^{-1} (H^e)^{\mathsf{T}} d^e \\
&\overset{(51.30)}{=} R^{-1} (R^{\mathsf{T}})^{-1} \begin{bmatrix} R^{\mathsf{T}} & 0 \end{bmatrix} Q^{\mathsf{T}} d^e \\
&\overset{(51.32)}{=} R^{-1} \begin{bmatrix} I_M & 0 \end{bmatrix} \begin{bmatrix} \bar{d} \\ \times \end{bmatrix} \\
&= R^{-1} \bar{d} \qquad (51.36)
\end{aligned}$$

We therefore arrive at the QR procedure listed in (51.37) for determining the ℓ_2-regularized solution, which involves solving a triangular system of equations.

QR method for minimizing ℓ_2-regularized least-squares risk (51.16).

given $\rho > 0$ and data $d = \mathrm{col}\{x(n)\}$, $H = \mathrm{blkrow}\{y_n^{\mathsf{T}}\}$;
construct $H^e = \begin{bmatrix} H \\ \sqrt{\rho N} I_M \end{bmatrix}$ and $d^e = \begin{bmatrix} d \\ 0_M \end{bmatrix}$;
perform the QR decomposition $H^e = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$;
apply Q^{T} to d^e and find $Q^{\mathsf{T}} d = \begin{bmatrix} \bar{d} \\ \times \end{bmatrix}$;
solve the triangular system of equations $R w_{\rm reg}^{\star} = \bar{d}$. (51.37)

51.3 ℓ_1-REGULARIZATION

In ℓ_1-regularization, we replace the empirical risk (51.1a) by the regularized version:

$$w_{\rm reg}^{\star} \triangleq \underset{w \in \mathbb{R}^M}{\mathrm{argmin}} \left\{ P_{\rm reg}(w) \triangleq \alpha \|w\|_1 + \frac{1}{N} \sum_{n=0}^{N-1} \left(x(n) - y_n^{\mathsf{T}} w\right)^2 \right\} \tag{51.38}$$

in terms of the ℓ_1-norm of w (i.e., the sum of its absolute entries), and where $\alpha > 0$ is the regularization factor; its value may or may not depend on N. In general, the value of α is independent of N. The variant with elastic-net

regularization solves instead

$$w^{\star}_{\rm reg} \;\triangleq\; \operatorname*{argmin}_{w\in\mathbb{R}^M} \left\{ P_{\rm reg}(w) = \alpha\|w\|_1 \;+\; \rho\|w\|^2 \;+\; \frac{1}{N}\sum_{n=0}^{N-1}\big(x(n) - y_n^{\mathsf{T}} w\big)^2 \right\} \tag{51.39}$$

with both $\alpha > 0$ and $\rho > 0$. We will discover in this section that ℓ_1-regularization leads to a sparse solution $w^{\star}_{\rm reg}$, i.e., to a solution with a few nonzero entries. In this way, for any observation vector y, the inner product calculation $\widehat{x} = y^{\mathsf{T}} w^{\star}_{\rm reg}$ ends up using only a few select entries from y due to the sparsity of $w^{\star}_{\rm reg}$. This means that ℓ_1-regularization performs a form of "dimensionality reduction." In particular, when some entries in y are correlated or redundant, the ℓ_1-solution will rely on one of them and ignore the others. Elastic-net regularization, on the other hand, inherits useful features from both ℓ_2 and ℓ_1-regularization. For example, it can handle situations involving more unknowns than measurements ($M > N$), and it also performs entry selection albeit in a less dramatic fashion than ℓ_1-regularization.

The following derivation extends Example 51.2 and provides a similar MAP interpretation for the ℓ_1-regularized empirical risk function (51.38).

Example 51.4 (**Interpretation in terms of a Laplacian prior on the model**) We collect N independent and identically-distributed observations $\{\boldsymbol{x}(n),\ y_n\}$, for $0 \le n \le N-1$, and assume that they satisfy the same linear model (51.17). The main difference is that we now assume that w is a realization of a random vector $\boldsymbol{w}$ whose entries $\{\boldsymbol{w}_m\}$ are independent of each other and arise from a Laplace distribution with zero mean and variance σ_w^2:

$$f_{\boldsymbol{w}_m}(w_m) \;=\; \frac{1}{\sqrt{2}\,\sigma_w}\exp\Big\{-\sqrt{2}|w_m|/\sigma_w\Big\} \tag{51.40}$$

We also assume that all observations $\{\boldsymbol{x}(n)\}$ are generated by the *same* realization w. We are again interested in estimating w. Using Bayes rule (3.39), we assess the conditional probability distribution of the model given the observations as follows:

$$\begin{aligned}
&f_{\boldsymbol{w}|\boldsymbol{x},\boldsymbol{y}}\Big(w|\{x(n),y_n\}\Big) \\
&\propto f_{\boldsymbol{x},\boldsymbol{y}|\boldsymbol{w}}\Big(\{x(n),y_n\}\,|\,w\Big)\,f_{\boldsymbol{w}}(w) \\
&= \left\{\prod_{n=0}^{N-1} f_{\boldsymbol{v}}\Big(x(n)-y_n^{\mathsf{T}}w\Big)\right\} f_{\boldsymbol{w}}(w) \\
&\propto \prod_{n=0}^{N-1}\exp\left\{-\frac{1}{2\sigma_v^2}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\} \times \prod_{m=1}^{M}\exp\Big\{-\sqrt{2}|w_m|/\sigma_w\Big\} \\
&= \exp\left\{-\frac{\sqrt{2}}{\sigma_w}\|w\|_1 - \frac{1}{2\sigma_v^2}\sum_{n=0}^{N-1}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\}
\end{aligned} \tag{51.41}$$

where the first and third lines replace the equality sign by proportionality constants. Consequently, we can now formulate a *maximum a-posteriori* (MAP) estimation prob-

lem to recover w by maximizing the above conditional density function as follows:

$$\begin{aligned} w^{\star}_{\rm reg} &\triangleq \operatorname*{argmax}_{w\in\mathbb{R}^M} \; f_{\boldsymbol{w}|\boldsymbol{x},\boldsymbol{h}}\Big(w|\{x(n),y_n\}\Big) \\ &= \operatorname*{argmin}_{w\in\mathbb{R}^M} \; \left\{\frac{\sqrt{2}}{\sigma_w}\|w\|_1 + \frac{1}{2\sigma_v^2}\sum_{n=0}^{N-1}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\} \\ &= \operatorname*{argmin}_{w\in\mathbb{R}^M} \; \frac{N}{2\sigma_v^2}\left\{\frac{2\sqrt{2}\sigma_v^2}{N\sigma_w}\|w\|_1 + \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\} \\ &= \operatorname*{argmin}_{w\in\mathbb{R}^M} \; \left\{\alpha\|w\|_1 + \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\} \end{aligned} \tag{51.42}$$

where we introduced $\alpha = 2\sqrt{2}\sigma_v^2/N\sigma_w$. We therefore recover the regularized empirical risk (51.38) with $q(w) = \alpha\|w\|_1$. This argument shows that ℓ_1-regularization helps ensure that the solution $w^{\star}_{\rm reg}$ is consistent with a prior Laplacian model on the distribution of $\boldsymbol{w}$.

We now explain how ℓ_1-regularization (or its extension in terms of elastic-net regularization) promotes solutions with smaller norm and alleviates the challenges of ill-conditioning, overfitting, and non-uniqueness of solutions.

Resolving non-uniqueness

The penalty term $\alpha\|w\|_1$ is only convex. The regularized risk function will have a unique minimizer $w^{\star}_{\rm reg}$ if the unregularized risk $P(w)$ happens to be strictly or strongly convex. For the least-squares case, the unregularized risk is strongly convex when $H^{\mathsf{T}}H > 0$. More generally, if this condition does not hold, then elastic-net regularization can be used and it will ensure a unique minimizer $w^{\star}_{\rm reg}$ because the resulting regularized risk in that case will become strongly-convex regardless of whether H is rank-deficient or not.

Promoting smaller solutions

It is seen from the regularized risk in (51.38)–(51.39) that larger values for α or ρ favor solutions $w^{\star}_{\rm reg}$ with smaller norms than would result when $\alpha = \rho = 0$. This is because the objective is to minimize the aggregate risk, and the regularization factors are influenced by $\alpha\|w\|_1$ and $\rho\|w\|^2$. This conclusion can be established more formally. If we set $q(w) = \alpha\|w\|_1$ for ℓ_1-regularization or $q(w) = \alpha\|w\|_1 + \rho\|w\|^2$ for elastic-net regularization, then it follows from the general result in Appendix 51.A that the following shrinkage property holds:

$$q(w^{\star}_{\rm reg}) \;\leq\; q(w^{\star}) \tag{51.43}$$

The result in the appendix holds for more general convex risks, $P(w)$, and is not limited to least-squares risks. It also holds for general convex regularization factors, $q(w)$, than ℓ_1 or elastic-net regularization. In other words, result (51.43) extends (51.24) to general convex risks and penalty terms.

Countering overfitting

Both ℓ_1 and elastic-net regularization help alleviate the danger of overfitting because they can also be shown to search for their solutions over reduced regions in space. This can be established more formally by verifying that minimizing a regularized least-squares problem of either forms (51.38)–(51.39) is equivalent to solving a *constrained* optimization problem of the following form:

$$w^{\star}_{\rm reg} \triangleq \underset{w\in\mathbb{R}^M}{\rm argmin} \left\{ \frac{1}{N}\sum_{n=0}^{N-1}(x(n)-y_n^{\mathsf{T}}w)^2 \right\}, \quad \text{subject to } q(w)\leq\tau \tag{51.44}$$

for some $\tau>0$ and using the appropriate regularization factor: $q(w)=\alpha\|w\|_1$ for ℓ_1-regularization and $q(w)=\alpha\|w\|_1+\rho\|w\|^2$ for elastic-net regularization. The equivalence between problems (51.38)–(51.39) and (51.44) is again established algebraically in Appendix 51.A by using the KKT multiplier arguments from Sec. 9.1.

Property (51.44) provides some intuition on how the choice of the penalty factor $q(w)$ defines the solution space. Figure 51.1 plots three contour curves in $2-$dimensional space corresponding to the level sets:

$$\|w\|_1=1, \quad \|w\|^2=1, \quad \|w\|_1+\|w\|^2=1 \tag{51.45}$$

It is seen from the figure that for ℓ_2-regularization, the search space for w is limited to a region delineated by a circular boundary. In comparison, the search space for ℓ_1-regularization is delineated by a rotated square boundary with sharp edges, while the search space for elastic-net regularization is midway between these two options. All three regions are obviously convex.

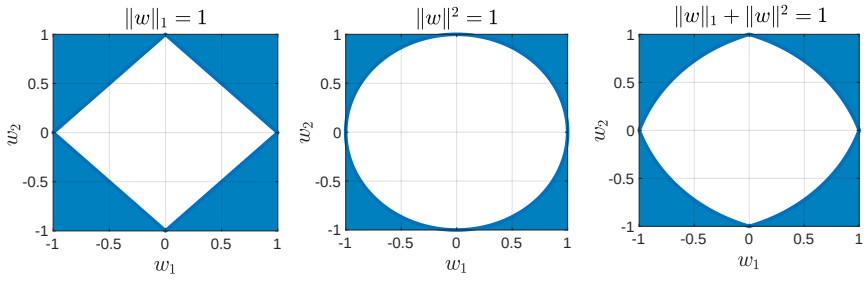

Figure 51.1 The figure illustrates the boundary curves corresponding to conditions (51.45) in $2-$dimensional space. The search space for the parameter, w, is limited to the inside of the regions delineated by these curves. Observe that in all three cases, the search domain is convex.

The particular shape for the boundary of the ℓ_1-region helps promote sparsity, i.e., it helps lead to solutions $w^{\star}_{\rm reg}$ with many zero entries. This is illustrated schematically in Fig. 51.2, which shows boundary curves corresponding to the regions $\|w\|_1\leq\tau$ and $\|w\|^2\leq\tau$, along with contour curves for the unregularized risk function, $P(w)$. The solution $w^{\star}_{\rm reg}$ occurs at the location where the contour

curves meet the boundary regions. It is seen, due to the corners that are present in the region $\|w\|_1 \leq \tau$, that the contour curves are more likely to touch this region at a corner point where some of the coordinates are zero. We will establish this conclusion more formally in the next section.

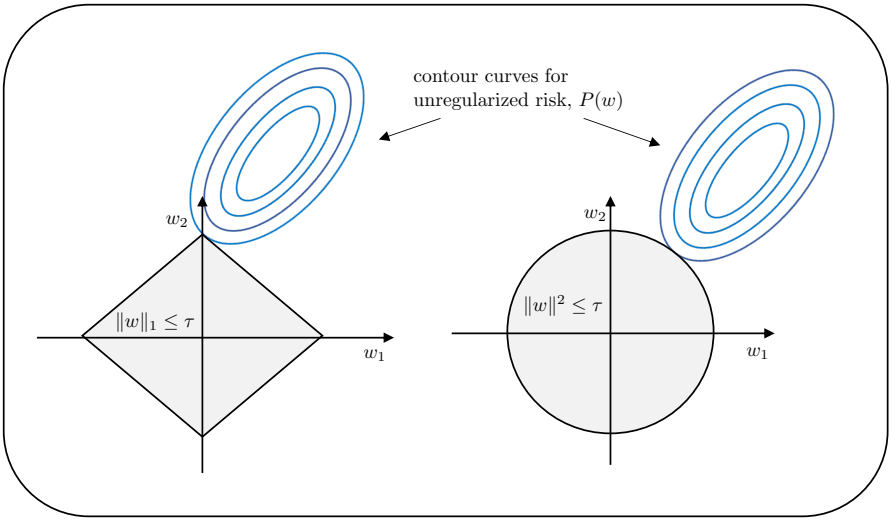

Figure 51.2 Boundary curves corresponding to the regions $\|w\|_1 \leq \tau$ and $\|w\|^2 \leq \tau$, along with contour curves for the unregularized risk function, $P(w)$.

51.4 SOFT THRESHOLDING

We are ready to examine the ability of ℓ_1-regularization to find sparse solution vectors, $w^{\star}_{\text{reg}}$. A sparse solution helps avoid over-fitting especially for large dimensional data (i.e., when M is large). This is because, when each observation vector y_t has many entries, a sparse $w^{\star}_{\text{reg}}$ assigns zero weights to those entries of y_t that are deemed "irrelevant." For this reason, we say that ℓ_1-regularization embodies an *automatic* selection capability into the solution by picking only entries from y_t that are most significant to the task of inferring $x(t)$.

For the benefit of the reader, we first review a useful result established earlier in Sec. 11.1.2 and which relies on the soft-thresholding function $\widehat{w} = \mathbb{T}_{\frac{\beta}{2}}(z)$. This function operates on the individual entries of its vector argument z to generate the corresponding entries of $\widehat{w}$. For each scalar x, the transformation $\mathbb{T}_{\frac{\beta}{2}}(x)$, with parameter $\beta \geq 0$, is defined as follows:

$$\mathbb{T}_{\frac{\beta}{2}}(x) \triangleq \begin{cases} x - \frac{\beta}{2}, & \text{if } x \geq \frac{\beta}{2} \\ 0, & \text{if } -\frac{\beta}{2} < x < \frac{\beta}{2} \\ x + \frac{\beta}{2}, & \text{if } x \leq -\frac{\beta}{2} \end{cases} \tag{51.46}$$

.

Lemma 51.1. (Soft-thresholding operation) *Given $z \in \mathbb{R}^M$, a constant $\beta \geq 0$, and a scalar ϕ, the solution to the optimization problem:*

$$\widehat{w} \triangleq \underset{w \in \mathbb{R}^M}{\operatorname{argmin}} \left\{ \beta \|w\|_1 + \|w - z\|^2 + \phi \right\} \tag{51.47}$$

is unique and given by

$$\widehat{w} = \mathbb{T}_{\frac{\beta}{2}}(z) \tag{51.48}$$

The soft-thresholding transformation $\mathbb{T}_{\frac{\beta}{2}}(z)$ helps promote sparse solutions $\widehat{w}$ (i.e., solutions with a few nonzero entries). This property is achieved in a measured manner since soft-thresholding sets to zero all entries of z whose magnitude is below the threshold value $\beta/2$, and reduces the size of the larger values by $\beta/2$. Figure 51.3 plots the function $\mathbb{T}_{\frac{\beta}{2}}(x)$ defined by (51.46). In summary, using the ℓ_1-penalty term in (51.47) results in a sparse solution $\widehat{w}$ that is "close" to the vector z.

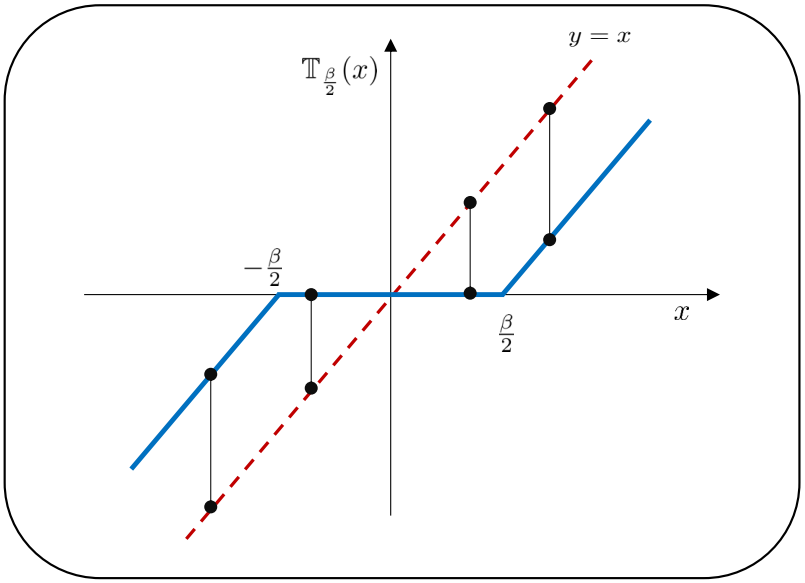

Figure 51.3 The soft-thresholding function, $\mathbb{T}_{\frac{\beta}{2}}(x)$, reduces the value of x gradually. Small values of x within the interval $[-\frac{\beta}{2}, \frac{\beta}{2}]$ are set to zero, while values of x outside this interval have their size reduced by an amount equal to $\beta/2$. The dotted segment represents the line $y = x$.

51.4.1 Orthogonal Data

Before studying the general case of arbitrary data matrices H, we consider first the special case when the "squared matrix" $H^{\mathsf{T}}H$ happens to be "orthogonal", namely, when H satisfies

$$H^{\mathsf{T}}H = \kappa^2 I_M, \quad \text{for some } \kappa^2 > 0 \tag{51.49}$$

Using this normalization condition, and the compact vector and matrix notation $\{d, H\}$ defined in (51.1c), we rewrite the unregularized and regularized risks in the form

$$\begin{aligned} P(w) &\triangleq \frac{1}{N}\sum_{n=0}^{N-1}\left(x(n) - y_n^{\mathsf{T}}w\right)^2 \\ &= \frac{1}{N}\|d - Hw\|^2 \\ &= \frac{1}{N}\left\{\|d\|^2 - 2d^{\mathsf{T}}Hw + \kappa^2\|w\|^2\right\} \end{aligned} \tag{51.50}$$

and

$$P_{\rm reg}(w) = \alpha\|w\|_1 \;+\; \frac{1}{N}\left\{\|d\|^2 - 2d^{\mathsf{T}}Hw + \kappa^2\|w\|^2\right\} \tag{51.51}$$

Note that both risks are strongly-convex since $\kappa^2 > 0$. Therefore, they each have a unique global minimizer, denoted by $w^\star$ and $w^\star_{\rm reg}$.

Lemma 51.2. (ℓ_1-regularized solution for orthogonal data) *Consider the ℓ_1-regularized problem (51.51) under the orthogonality condition (51.49). The solution is unique and given by*

$$w^\star_{\rm reg} \;=\; \mathbb{T}_{\frac{\alpha N}{2\kappa^2}}\left(w^\star\right) \tag{51.52}$$

where $w^\star = \frac{1}{\kappa^2}H^{\mathsf{T}}d$ is the minimizer for the unregularized risk (51.50).

Proof: We employ a completion-of-squares argument to write (51.51) as

$$P_{\rm reg}(w) = \alpha\|w\|_1 \;+\; \frac{\kappa^2}{N}\left\{\|w\|^2 - \frac{2}{\kappa^2}d^{\mathsf{T}}Hw + \frac{1}{\kappa^2}\|d\|^2\right\} nn \tag{51.53}$$

$$\begin{aligned} &= \alpha\|w\|_1 \;+\; \frac{\kappa^2}{N}\left\{\left\|w - \frac{1}{\kappa^2}H^{\mathsf{T}}d\right\|^2 + \frac{1}{\kappa^2}\|d\|^2 - \frac{1}{\kappa^4}\|H^{\mathsf{T}}d\|^2\right\} \\ &\propto \beta\|w\|_1 \;+\; \|w - z\|^2 \;+\; \phi \end{aligned} \tag{51.54}$$

where $\propto$ is the proportionality symbol, while the scalars $\{\beta, \phi\}$ and the column vector $z \in \mathbb{R}^M$ are defined by

$$\beta \;\triangleq\; \frac{\alpha N}{\kappa^2} > 0 \tag{51.55a}$$

$$z \;\triangleq\; \frac{1}{\kappa^2}H^{\mathsf{T}}d \;=\; w^\star \tag{51.55b}$$

$$\phi \;\triangleq\; \frac{1}{\kappa^2}\|d\|^2 - \frac{1}{\kappa^4}\|H^{\mathsf{T}}d\|^2 \tag{51.55c}$$

Observe that z agrees with the minimizer, $w^\star$, for the unregularized problem under condition $H^{\mathsf{T}}H = \kappa^2 I$. Minimization of the empirical risk (51.54) is now of the same form as problem (51.47). Therefore, we deduce that the minimizer to (51.49) under the orthogonality condition (51.49) is given by (51.52).

■

Observe how construction (51.52) applies soft-thresholding to $w^\star$ with the threshold defined by $\alpha N/2\kappa^2$; this value (and, hence, sparsity) increases with α.

51.4.2 LASSO or Basis Pursuit Denoising

More generally, for data matrices H that do not satisfy the orthogonality condition (51.49), we can derive a similar expression for $w^\star_{\rm reg}$ involving a soft-thresholding operation, albeit one where $w^\star$ is replaced by another vector defined in terms of a *dual* variable — see expression (51.61) further ahead. We will derive the result under both ℓ_1 and elastic-net regularization.

Thus, consider the regularized least-squares problem:

$$w^\star_{\rm reg} = \operatorname*{argmin}_{w\in\mathbb{R}^M} \left\{ P_{\rm reg}(w) \;\triangleq\; q(w) \;+\; \frac{1}{N}\sum_{n=0}^{N-1} \big(x(n) - y_n^{\mathsf{T}} w\big)^2 \right\} \tag{51.56a}$$

where the regularization factor has the form

$$q(w) = \alpha\|w\|_1 + \rho\|w\|^2, \quad \alpha > 0, \quad \rho \geq 0 \tag{51.56b}$$

When $\rho = 0$ we have pure ℓ_1-regularization. Problem (51.56a)–(51.56b) is known as LASSO, where the acronym stands for "*least absolute shrinkage and selection operator*." The shrinkage feature is because the resulting solution $w^\star_{\rm reg}$ will end up satisfying property (51.43). The selection feature is because the same solution will be sparse with generally few nonzero entries. Problem (51.56a) is also known as the *basis pursuit denoising problem*; this is because it seeks a sparse representation for the vector d in terms of the columns of H.

Unfortunately, when H is not orthogonal, a closed-form expression for the solution $w^\star_{\rm reg}$ is not possible any longer. For this reason, the LASSO problem (51.56a) is usually solved iteratively by means of subgradient or proximal gradient iterations, with or without stochastic sampling of data, as was already shown earlier in several instances including in Examples 14.1, 15.3, and 16.12; the latter example describes a *stochastic proximal* gradient implementation that relies on instantaneous gradient approximations and which we reproduce here illustration purposes.

Stochastic proximal gradient algorithm for LASSO problem (51.56a)

given dataset $\{x(m), y_m\}_{m=0}^{N-1}$;
start from an arbitrary initial condition, $\boldsymbol{w}_{-1}$.
repeat until convergence over $n \geq 0$:
  select at random a sample $(\boldsymbol{x}(n), \boldsymbol{y}_n)$ at iteration n;
  $\boldsymbol{z}_n = (1 - 2\mu\rho)\boldsymbol{w}_{n-1} + 2\mu\boldsymbol{y}_n(\boldsymbol{x}(n) - \boldsymbol{y}_n^{\mathsf{T}}\boldsymbol{w}_{n-1})$
  $\boldsymbol{w}_n = \mathbb{T}_{\mu\alpha}(\boldsymbol{z}_n)$
end
return $w^\star \leftarrow \boldsymbol{w}_n$.

(51.57)

Other implementations are of course possible. For instance, Example 15.3 describes a full-batch implementation leading to the iterated soft-thresholding algorithm (ISTA):

$$\begin{cases} z_n = (1 - 2\mu\rho)w_{n-1} + \dfrac{2\mu}{N}\displaystyle\sum_{m=0}^{N-1}(x(m) - y_m^{\mathsf{T}}w_{n-1}) \\ w_n = \mathbb{T}_{\mu\alpha}(z_n) \end{cases} \tag{51.58}$$

Numerical solutions of the LASSO optimization problem based on the use of convex optimization packages are also possible. The derivation in this section is meant to highlight some properties of the exact solution, such as showing that it continues to have a soft-thresholding form. To do so, we will follow a duality argument.

Expression for LASSO solution

Using the vector notation $\{d, H\}$, problem (51.56a) can be recast as

$$w_{\text{reg}}^\star = \underset{w\in\mathbb{R}^M}{\text{argmin}} \quad \left\{ q(w) \;+\; \frac{1}{N}\|d - Hw\|^2 \right\} \tag{51.59}$$

or, equivalently, in terms of an auxiliary variable transformation that introduces a constraint:

$$\begin{aligned} (w_{\text{reg}}^\star, z^\star) = \quad & \underset{w,z}{\text{argmin}} \quad \left\{ q(w) \;+\; \frac{1}{N}\|d - z\|^2 \right\} \\ & \text{subject to } z = Hw \end{aligned} \tag{51.60}$$

where we introduced $z \in \mathbb{R}^N$; it depends linearly on w. The risk function in statement (51.60) is convex over w and z. We therefore have a convex optimization problem with a linear equality constraint. This type of formulation is a special case of problem (9.1), involving convex costs subject to convex inequality and equality constraints, and which we studied in Sec. 9.1. The results from that section show that strong duality holds for problem (51.60). This means that we can

learn about the solution $w^{\star}_{\rm reg}$ by using duality arguments to establish the next theorem for both cases of $\rho \neq 0$ and $\rho = 0$; the proof appears in Appendix 51.B.

Theorem 51.1. (Expression for LASSO solution) *Consider the regularized problem (51.56a)–(51.56b). The solution is unique and admits the following representation:*

$$w^{\star}_{\rm reg} \;=\; \frac{1}{2\rho}\,\mathbb{T}_{\alpha}\left(H^{\mathsf{T}}\lambda^{o}\right) \tag{51.61}$$

where $\lambda^{o} \in \mathbb{R}^{N}$ is determined as follows:

(a) **(elastic-net regularization**, $\rho \neq 0$**)**: *λ^{o} is the unique maximum of the following strongly-concave function:*

$$\lambda^{o} = \underset{\lambda\in\mathbb{R}^{N}}{\rm argmax}\;\left\{\lambda^{\mathsf{T}}d - \frac{N}{4}\|\lambda\|^{2} - \frac{1}{4\rho}\left\|\mathbb{T}_{\alpha}\left(H^{\mathsf{T}}\lambda\right)\right\|^{2}\right\} \tag{51.62}$$

(b) **(ℓ_1-regularization**, $\rho = 0$**)**: *λ^{o} is the unique projection of the vector $\frac{2}{N}d$ onto the set of vectors λ satisfying $\|H^{\mathsf{T}}\lambda\|_{\infty} \leq \alpha$:*

$$\lambda^{o} = \underset{\lambda\in\mathbb{R}^{N}}{\rm argmin}\;\left\|\lambda - \frac{2}{N}d\right\|^{2}, \quad \text{subject to } \|H^{\mathsf{T}}\lambda\|_{\infty} \leq \alpha \tag{51.63}$$

Comparing expression (51.61) with (51.52) for "orthogonal" data matrices, we note that the argument of the soft-thresholding function is now defined in terms of a dual variable λ^{o} and not in terms of the unregularized solution, $w^{\star}$. Moreover, the threshold in $\mathbb{T}_{\alpha}(\cdot)$ increases with α so that more sparse models are expected for larger α. Clearly, solving the LASSO problem via (51.61) is not simpler than solving the original optimization problem (51.56a) because we still need to determine λ^{o} in (51.62) or (51.63). The usefulness of result (51.61) is that it provides a representation for the solution in a manner similar to (51.52) and helps illustrate the sparsity properties of the resulting $w^{\star}_{\rm reg}$. The parameters α and ρ define the degree of regularization: larger values tend to promote smaller (in norm) and more sparse solutions. One useful way to select these parameters is the cross validation technique described later in Sec. 61.3.

Example 51.5 **(Comparing different regularized solutions)** In this example we compare numerically the behavior of ℓ_2, ℓ_1, and elastic-net regularization solutions. First, however, we need to show how to approximate the regularized solution to (51.56a)–(51.56b). We already know that we can employ a stochastic subgradient algorithm for this purpose to arrive at good approximations for $w^{\star}_{\rm reg}$. Under elastic-net regularization, the recursion would start from some random initial guess, denoted by w_{-1}, and then iterate as follows:

$$w_n = (1-2\mu\rho)w_{n-1} - \mu\,\alpha\,\mathrm{sign}(w_{n-1}) + 2\mu y_n(x(n) - y_n^{\mathsf{T}}w_{n-1}), \quad n \geq 0 \tag{51.64}$$

where μ is a small step-size parameter and the notation w_n denotes the approximation for the regularized solution at iteration n. The sign function, when applied to a vector argument, returns a vector with entries equal to ± 1 depending on the signs of the individual entries of w_{n-1}: $+1$ for nonnegative entries and -1 for negative entries. The

algorithm is run *multiple* times over the training data $\{x(n), y_n\}$, with the data being randomly reshuffled at the beginning of each epoch, namely,

(a) At the start of each epoch, the data $\{x(n), y_n\}$ is randomly reshuffled so that each epoch runs over the same dataset albeit in a different random order.

(b) The initial condition for the epoch of index k is the iterate value that was obtained at the end of the previous epoch.

The iterate that is obtained at the end of the last epoch is the one that is taken to be the approximation for $w^{\star}_{\rm reg}$.

Iteration (51.64) applies to both cases of ℓ_1-regularization (by setting $\rho = 0$) and elastic-net regularization when both α and ρ are positive. Although we already have a closed-form solution for the ℓ_2-regularized solution via expression (51.21), or can even arrive at it by means of the recursive least-squares (RLS) algorithm (50.123), the same stochastic recursion (51.64) can be used to approximate the ℓ_2-regularized solution as well by setting $\alpha = 0$; the recursion leads to a computationally simpler algorithm than RLS albeit at a slower convergence rate.

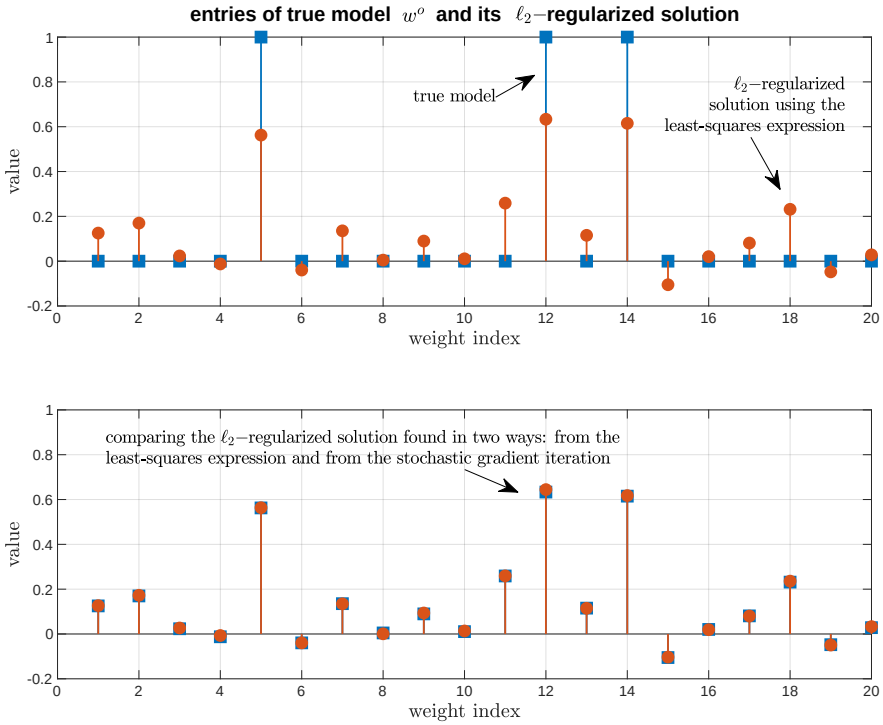

Figure 51.4 The top plot shows the true model w^o with three nonzero entries at value one while all other entries are at zero. The top plot also shows the ℓ_2-regularized solution, $w^{\star}_{\rm reg}$, that is obtained by using the least-squares expression (51.21). The bottom plot compares the solutions that are obtained from the least-squares expression (51.21) and from the stochastic recursion (51.64) using 20 runs over the data, $\alpha = 0$, and $\mu = 0.0001$. It is seen that recursion (51.64) is able to learn the ℓ_2-regularized solution well.

We use the stochastic construction (51.64) to illustrate the behavior of the different

regularization modes, by considering the following numerical example. We generate $N = 4000$ random data points $\{\boldsymbol{x}(n), \boldsymbol{y}_n\}$ related through the linear model:

$$\boldsymbol{x}(n) = \boldsymbol{y}_n^{\mathsf{T}} w^o + \boldsymbol{v}(n) \tag{51.65}$$

where $\boldsymbol{v}(n)$ is while Gaussian noise with variance $\sigma_v^2 = 0.01$, and each observation vector has dimension $M = 20$. We generate a sparse true model w^o consisting of three randomly-chosen entries set to one, while all other entries of w^o are set to zero. Figures 51.4 and 51.5 illustrate the results that follow from using 20 runs over the data with $\mu = 0.0001$, $\alpha = 5$, and $\rho = 2$. It is seen in the lower plot from the first figure that the stochastic recursion (51.64) converges to a good approximation for the actual least-squares solution from (51.21). The middle plot of the second figure illustrates the sparsity property of the ℓ_1-regularized solution.

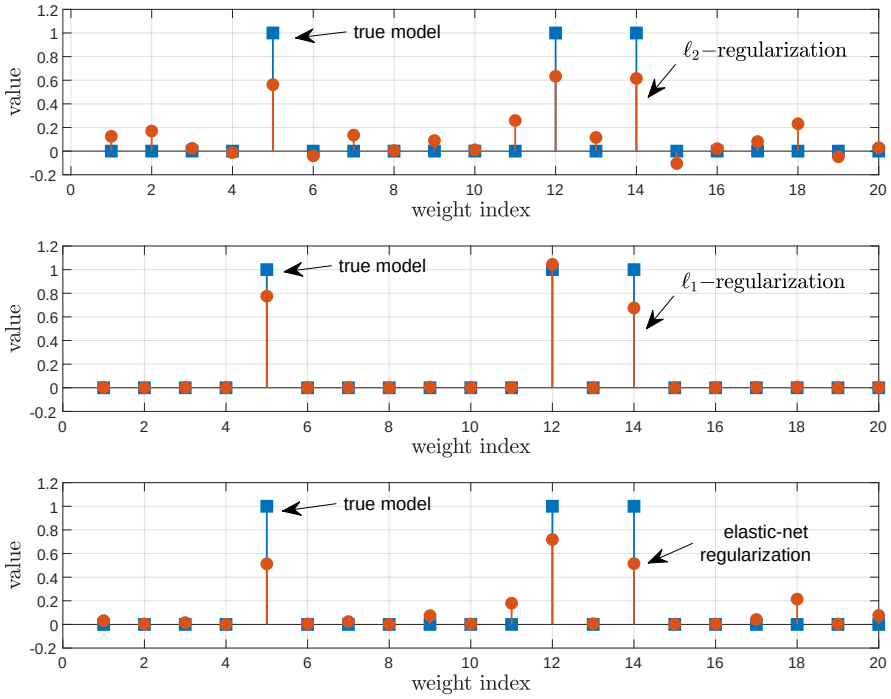

Figure 51.5 All three plots show the true model w^o with three nonzero entries at value one while all other entries are at zero. In each case, the true model is compared against the ℓ_2-regularized solution (*top plot*), the ℓ_1-regularized solution using $\alpha = 5$ (*middle plot*), and the elastic-net regularized solution using $\alpha = 5$ and $\rho = 2$ (*bottom plot*). All these regularized solutions are obtained by using the stochastic (sub)gradient recursion (51.64) using 20 runs over the data and $\mu = 0.0001$. The middle plot illustrates how ℓ_1-regularization leads to a sparse solution, while the elastic-net regularized solution has slightly more non-zero entries.

51.5 COMMENTARIES AND DISCUSSION

Tikhonov regularization. The regularized least-squares problem (51.16) and its solution (51.21) were proposed by the Russian mathematician **Andrey Tikhonov (1906–1993)** in the publication by Tikhonov (1963) on ill-posed problems — see also the text by Tikhonov and Arsenin (1977). This form of regularization is nowadays very popular and is known as *Tikhonov regularization*. Tikhonov's formulation was general and applicable to infinite-dimensional operators and not only to finite-dimensional least-squares problems. His work was aimed at solving integral equations of the first-kind, also known as Fredholm integral equations, which deal with the problem of determining a function solution $x(t)$ to an integral equation of the following form:

$$\int_a^b A(s,t)x(t)dt = b(s) \tag{51.66}$$

for a given kernel function, $A(s,t)$, and another function $b(s)$. These integral equations can be ill-conditioned and can admit multiple solutions. The analogy with linear systems of equations of the form $Ax = b$ becomes apparent if we employ the operator notation to rewrite the integral equation in the form $\mathcal{A}x = b$, in terms of some infinite-dimensional operator $\mathcal{A}$. It turns out that both Phillips (1962) and Tikhonov (1963) proposed using ℓ_2-regularization to counter ill-conditioning for Fredholm integral equations, which is why this type of regularization is also referred to as the Phillips–Tikhonov or Tikhonov–Phillips regularization. The same technique also appeared in Hoerl (1962), albeit for finite-dimensional operators (i.e., for matrices) in the context of least-squares problems. This latter work was motivated by the earlier contribution on ridge analysis from Hoerl (1959) — see also Hoerl and Kennard (1970) and the review by Hoerl (1985). It is for this reason that ℓ_2-regularization is also known as ridge regression in the statistics literature. Useful overviews on the role of Tikhonov regularization in the solution of linear systems of equations and least-squares problems appear in the survey article by Neumaier (1998) and in the texts by Golub and Van Loan (1996), Bjorck (1996), and Hansen (1997). More information on regularization in general can be found in the texts by Wahba (1990) and Engl, Hanke, and Neubauer (1996).

LASSO and basis pursuit denoising. In Examples 51.2 and 51.4 we showed that regularization in the least-squares case corresponds to associating a prior distribution with the sought-after parameter, $\boldsymbol{w}$ (now treated as a random quantity). A Gaussian prior leads to ℓ_2-regularization, while a Laplacian prior leads to ℓ_1-regularization as noted by Tibshirani (1996b). We showed in the body of the chapter that ℓ_1-regularization leads to sparse solutions. However, it has been observed in practice that it tends to retain more nonzero entries than necessary in the solution vector and, moreover, if several entries in the observation space are strongly correlated, the solution vector will tend to keep one of them and discard the others — see Zou and Hastie (2005). Elastic-net regularization, on the other hand, combines ℓ_1 and ℓ_2-penalty terms and inherits some of their advantages: it promotes sparsity without totally discarding highly correlated observations. This form of regularization was proposed by Zou and Hastie (2005); examination of some of its properties appears in this reference as well as in the text by Hastie, Tibshirani, and Friedman (2009) and in De Mol, De Vito, and Rosasco (2009).

Given data $\{x(n), y_n \in \mathbb{R}^M\}$, the pure ℓ_1-regularization formulation solves

$$w^\star_{\text{reg}} = \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{ \alpha\|w\|_1 \;+\; \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n) - y_n^{\mathsf{T}}w\Big)^2 \right\} \tag{51.67}$$

where $\alpha > 0$ is the regularization parameter. We explained in the chapter that this

problem is equivalent to solving

$$w^{\star}_{\rm reg} = \underset{w\in\mathbb{R}^M}{\rm argmin}\ \|d-Hw\|^2, \quad \text{subject to } \alpha\|w\|_1 \le \tau \tag{51.68}$$

for some $\tau > 0$. Problems of this type were first proposed by Santosa and Symes (1986) and later by Tibshirani (1996b); the latter reference uses the acronym LASSO for such problems. A similar problem was studied by Chen, Donoho, and Saunders (1998,2001) under the name *basis pursuit denoising*. They examined instead the reverse formulation:

$$w^{\star}_{\rm reg} = \underset{w\in\mathbb{R}^M}{\rm argmin}\ \|w\|_1, \quad \text{subject to } \|d-Hw\|^2 \le \epsilon \tag{51.69}$$

for some small $\epsilon > 0$. This formulation was motivated by the earlier work in Chen and Donoho (1994) on standard *basis pursuit*. In this latter problem, the objective is to seek a sparse representation for a signal vector d from an overcomplete basis H, namely, to solve — see Prob. 51.7:

$$\min_{w\in\mathbb{R}^M} \|w\|_1, \quad \text{subject to } d = Hw \tag{51.70}$$

All three formulations (51.67), (51.68), and (51.69) are equivalent to each other for suitable choices of the parameters $\{\alpha, \tau, \epsilon\}$ — see Prob. 51.7. The contributions by Tibshirani (1996b) and Chen, Donoho, and Saunders (1998,2001) generated renewed interest in ℓ_1-regularized problems in the statistics, machine learning, and signal processing literature. These types of problems have an older history, especially in the field of geophysics. For example, a problem of the same form as (51.67) was used in the deconvolution of seismic signals by Santosa and Symes (1986). Their work was motivated by the earlier contributions by Claerbout and Muir (1973) and Taylor, Banks, and McCoy (1979). Using our notation, these last two references consider optimization problems of the following form (compare with (51.67)):

$$w^{\star}_{\rm reg} = \underset{w\in\mathbb{R}^M}{\rm argmin}\ \left\{ \alpha\|w\|_1 \ +\ \frac{1}{N}\sum_{n=0}^{N-1} |x(n) - y_n^{\mathsf{T}} w| \right\} \tag{51.71}$$

where the sum of the absolute residuals (rightmost term) is used in place of the sum of their squared values, as is the case in (51.67). Both formulations employ an ℓ_1-penalty term. One of the earliest recognitions that ℓ_1-regularization promotes sparsity appears in the article by Santosa and Symes (1986, p. 1308), where it is stated that the use of the ℓ_1-penalty term *"has the effect of constructing a solution which has the least number of nonzero components."* Arguments and derivations in support of the sparsity-promoting property of the ℓ_1-penalty appear in Levy and Fullagar (1981), Oldenburg, Scheuer, and Levy (1983), and also in Santosa and Symes (1986, Sec. 2). In their formulation of the deconvolution problem, Santosa and Symes (1986) proposed replacing (51.71) by the same problem (51.67) using the sum of squared residuals — see their expressions (1.16) and (5.1).

It is useful to note that design problems involving ℓ_1-measures of performance have also been pursued in the control field, starting from the mid 1980s. The primary motivation there for the use of the ℓ_1-norm has been to design control laws that minimize the effect of persistent *bounded* disturbances on the output of the system. Among the earliest references that promoted this approach are the works by Vidyasagar (1986) and Dahleh and Pearson (1986,1987). A thorough treatment of the subject matter, along with an extensive bibliography, appears in the text by Dahleh and Diaz-Bobillo (1995).

Robust least-squares designs. Given an $N \times M$ data matrix H, an $N \times 1$ target vector d, an $N \times N$ positive-definite weighting matrix R, and an $M \times M$ positive-definite regularization matrix Π, the solution to the following regularized weighted least-squares

problem:

$$w^\star \triangleq \operatorname*{argmin}_{w\in\mathbb{R}^M} \Big\{ w^{\mathsf{T}}\Pi w + (d-Hw)^{\mathsf{T}} R (d-Hw) \Big\} \tag{51.72}$$

is unique and given by

$$w^\star = (\Pi + H^{\mathsf{T}} R H)^{-1} H^{\mathsf{T}} R\, d \tag{51.73}$$

When the data $\{d, H\}$ are subject to uncertainties, the performance of this solution can deteriorate appreciably. Assume that the actual data matrix that generated the target signal d is $H+\delta H$ and not H, for some small perturbation δH. Then, the above solution $w^\star$, which is designed based on knowledge of the nominal value H, does not take into account the presence of the perturbations in the data. One way to address this problem is to formulate a robust version of the least-squares problem as follows:

$$w^{\mathrm{rob}} \triangleq \tag{51.74a}$$

$$\operatorname*{argmin}_{w\in\mathbb{R}^M} \max_{\{\delta H, \delta d\}} \left\{ w^{\mathsf{T}}\Pi w + \Big((d+\delta d) - (H+\delta H)w\Big)^{\mathsf{T}} R \Big((d+\delta d) - (H+\delta H)w\Big) \right\}$$

where $\{\delta d, \delta H\}$ represent the unknown perturbations that are assumed to be modeled as follows:

$$\begin{bmatrix} \delta d & \delta H \end{bmatrix} = P \Delta \begin{bmatrix} e_d & E_H \end{bmatrix} \tag{51.74b}$$

where Δ is an arbitrary contraction matrix satisfying $\|\Delta\| \leq 1$ and $\{P, e_d, E_H\}$ are known quantities of appropriate dimensions, e.g., e_d is a column vector. The matrix P is meant to constrain the perturbations to its range space. Problem (51.74a) can be interpreted as a constrained two-game problem, with the designer trying to select an estimate w^{rob} that minimizes the cost while the opponent $\{\delta d, \delta H\}$ tries to maximize the same cost. It turns out that the solution to (51.74a) has the form of a *regularized* least-squares solution albeit one with modified $\{\Pi, R\}$ matrices, as indicated by the following result.

Robust regularized least-squares (Sayed, Nascimento, and Cipparrone (2002)). *Problem (51.74a)–(51.74b) has a unique solution given by*

$$w^{\mathrm{rob}} = \Big(\widehat{\Pi} + H^{\mathsf{T}}\widehat{R}H\Big)^{-1}\Big(H^{\mathsf{T}}\widehat{R}\,d + \widehat{\beta}E_H^{\mathsf{T}}e_d\Big) \tag{51.75a}$$

where $\{\widehat{\Pi}, \widehat{R}\}$ are obtained from $\{\Pi, R\}$ as follows:

$$\widehat{\Pi} = \Pi + \widehat{\beta}E_H^{\mathsf{T}}E_H \tag{51.75b}$$

$$\widehat{R} = R + RP(\widehat{\beta}I_M - P^{\mathsf{T}}RP)^{\dagger}P^{\mathsf{T}}R \tag{51.75c}$$

where the notation $\dagger$ refers to the pseudo inverse of its matrix argument, and the scalar $\widehat{\beta}$ is determined by solving

$$\widehat{\beta} = \operatorname*{argmin}_{\beta \geq \|P^{\mathsf{T}}RP\|} G(\beta) \tag{51.75d}$$

where the function $G(\beta)$ is defined as follows:

$$G(\beta) \triangleq \|w(\beta)\|^2_{\Pi(\beta)} + \|d - Hw(\beta)\|^2_{R(\beta)} + \beta\|e_d - E_H w(\beta)\|^2 \tag{51.76}$$

where the notation $\|a\|_X^2$ *stands for* $a^{\mathsf{T}}Xa$ *and*

$$R(\beta) = R + RP(\beta I - P^{\mathsf{T}}RP)^{\dagger}P^{\mathsf{T}}R \tag{51.77a}$$

$$\Pi(\beta) = \Pi + \beta E_H^{\mathsf{T}}E_H \tag{51.77b}$$

$$w(\beta) = \left(\widehat{\Pi}(\beta) + H^{\mathsf{T}}R(\beta)H\right)^{-1}\left(H^{\mathsf{T}}R(\beta)\,d + \beta E_H^{\mathsf{T}}e_d\right) \tag{51.77c}$$

We denote the lower bound on β by $\beta_\ell = \|P^{\mathsf{T}}RP\|$. Compared with the solution (51.73) to the original regularized least-squares problem, we observe that the expression for $w^{\rm rob}$ is distinct in some important ways:

(a) First, the weighting matrices $\{\Pi, R\}$ are replaced by corrected versions $\{\widehat{\Pi}, \widehat{R}\}$. These corrections are defined in terms of a scalar $\widehat{\beta}$, which is obtained by minimizing $G(\beta)$ over the semi-open interval $[\beta_\ell, \infty)$.

(b) It was shown by Sayed and Chen (2002) and Sayed, Nascimento, and Cipparrone (2002) that the function $G(\beta)$ has a unique global minimum (and no local minima) over the interval $[\beta_\ell, \infty)$. This means that the determination of $\widehat{\beta}$ can be pursued by standard search procedures without worrying about convergence to undesired local minima. Extensive experiments suggest that setting $\widehat{\beta} = \lambda\beta_\ell$ (a scaled multiple of the lower bound for some positive λ chosen by the designer) is generally sufficient.

(c) The right-hand side of (51.75a) contains an additional term $\widehat{\beta}E_H^{\mathsf{T}}e_d$. The expression for $w^{\rm rob}$ can be viewed as the solution to the following extended problem

$$w^{\rm rob} = \operatorname*{argmin}_{w\in\mathbb{R}^M}\left\{\begin{bmatrix} 1 & w^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}\widehat{\beta}\|e_d\|^2 & -\widehat{\beta}e_d^{\mathsf{T}}E_H \\ -\widehat{\beta}E_H^{\mathsf{T}}e_d & \widehat{\Pi}\end{bmatrix}\begin{bmatrix}1\\ w\end{bmatrix} + \|d - Hw\|_{\widehat{R}}^2\right\} \tag{51.78}$$

(d) For values $\widehat{\beta} > \beta_\ell$, the pseudo-inverse operation can be replaced by standard matrix inversion and it holds that

$$\widehat{R}^{-1} = R^{-1} - \widehat{\beta}^{-1}PP^{\mathsf{T}} \tag{51.79}$$

Other robust variations of least-squares are possible. For example, model (51.74b) for the perturbations can be replaced by one of the form

$$\|\delta H\| \le \eta, \quad \|\delta d\| \le \eta_d \tag{51.80}$$

where the uncertainties are instead assumed to lie within bounded regions determined by the positive scalars $\{\eta, \eta_d\}$. The solution has a similar structure and is described in Chandrasekaran *et al.* (1997,1998) and Sayed, Nascimento, and Cipparrone (2002). A convex optimization approach is described in El Ghaoui and Lebret (1997). Other variations and geometric arguments are described in Sayed, Nascimento, and Chandrasekaran (1998) — see also Probs. 51.19–51.21.

PROBLEMS

51.1 Consider the least-squares problem (51.1a) with a rank deficient H:

$$H = \begin{bmatrix}1 & 2\\ 1 & 2\\ 0 & 0\end{bmatrix}, \quad d = \begin{bmatrix}+1\\ +1\\ -1\end{bmatrix}$$

(a) Verify that all solutions to the normal equations take the form $w^\star = \mathrm{col}\{1-2b, b\}$ for any $b \in \mathbb{R}$.
(b) Verify that all vectors in the nullspace of $H^{\mathsf{T}}H$ take the form $p = \mathrm{col}\{-2b, b\}$.
(c) Verify that the following are two valid solutions:

$$w_1^\star = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad w_2^\star = \begin{bmatrix} -3 \\ 2 \end{bmatrix}$$

(d) Consider the test vector $y_t = \mathrm{col}\{2, 2\}$. Compute the $\widehat{x}(t)$ that result from both solutions. *Remark.* Observe that the predictions have opposite signs, which is undesirable in applications where the sign of $\widehat{x}(t)$ is used to perform classification.

51.2 Consider the least-squares problem (51.1a) with an ill-conditioned matrix H:

$$H = \begin{bmatrix} 1 & & \\ & -1 & \\ & & \sqrt{\epsilon} \end{bmatrix}, \quad d = \begin{bmatrix} +1 \\ -1 \\ +1 \end{bmatrix} \tag{51.81}$$

where $\epsilon > 0$ is a small number, and the entries of d are binary variables of the type $x(n) = \pm 1$.
(a) What is the condition number of $H^{\mathsf{T}}H$?
(b) Determine the solution $w^\star$ to the normal equations.
(c) Consider the two observation vectors

$$y_1 = \mathrm{col}\left\{10, 10, 10^{-6}\right\}, \quad y_2 = \mathrm{col}\left\{10, 10, -10^{-6}\right\}$$

where their trailing entries have small size and differ in sign. Predict their target signals $\widehat{x}(1)$ and $\widehat{x}(2)$. *Remark.* Observe how $\widehat{x}_2$ can become negative for small enough ϵ while $\widehat{x}_1$ is always positive. If the sign of $\widehat{x}$ is used to classify the observation vector y, then the vectors $\{y_1, y_2\}$, despite being very close to each other in Euclidean space, will end up being assigned to different classes.

51.3 Let $w^\star$ and $w_{\mathrm{reg}}^\star$ denote solutions to the unregularized and regularized least-squares risks (51.1a) and (51.16), respectively.
(a) Show that $w_{\mathrm{reg}}^\star = (\rho N I + H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}Hw^\star$.
(b) Introduce the eigen-decomposition $H^{\mathsf{T}}H = U\Lambda U^{\mathsf{T}}$, where U is $M \times M$ orthogonal and Λ is diagonal with nonnegative entries $\{\lambda(m)\}$. Let $\bar{w}_{\mathrm{reg}} = U^{\mathsf{T}}w_{\mathrm{reg}}^\star$ and $\bar{w} = U^{\mathsf{T}}w^\star$ and denote their individual entries by $\{\bar{w}_{\mathrm{reg}}(m),\ \bar{w}(m)\}$. Verify that

$$\bar{w}_{\mathrm{reg}}(m) \ = \ \left(\frac{\lambda(m)}{\rho N + \lambda(m)}\right) \bar{w}(m), \quad m = 1, 2, \ldots, M$$

Conclude that $\|w_{\mathrm{reg}}^\star\|^2 < \|w^\star\|^2$.

51.4 Refer to the minimizers $\{w^\star, w_{\mathrm{reg}}^\star\}$ for the unregularized and regularized least-squares problems.
(a) Show that $P_{\mathrm{reg}}(w_{\mathrm{reg}}^\star) - P(w^\star) \ = \ \rho\,(w^\star)^{\mathsf{T}}w_{\mathrm{reg}}^\star$.
(b) Introduce the same transformations $\{\bar{w}, \bar{w}_{\mathrm{reg}}\}$ from Prob. 51.3 and conclude that

$$P_{\mathrm{reg}}(w_{\mathrm{reg}}^\star) \ - \ P(w^\star) = \sum_{m=1}^{M} \left(\frac{\rho\lambda(m)}{\rho N + \lambda(m)}\right) |\bar{w}(m)|^2$$

(c) Since generally at least one $\lambda(m) \neq 0$ and $w^\star \neq 0$, conclude that $P_{\mathrm{reg}}(w_{\mathrm{reg}}^\star) > P(w^\star)$.
(d) Verify that the function $f(\rho) = \rho\lambda/(\rho N + \lambda)$ is non-decreasing in ρ. Conclude that the bias increases with ρ.

51.5 We re-examine the result of Prob. 50.11 for the case of ℓ_2-regularized least-squares (or ridge regression). Thus, refer again to the stochastic model (50.88) where $\boldsymbol{v}$ has covariance matrix $\sigma_v^2 I_N$ but is not necessarily Gaussian. Introduce the mean-square error risk, $P(w) = \mathbb{E}\,\|\boldsymbol{d} - Hw\|^2$, where the expectation is over the source of randomness

in $\boldsymbol{d}$. Verify that the ℓ_2-regularized least-squares solution $\boldsymbol{w}^\star_{\rm reg}$ given by (51.21) leads to the following average excess risk expression:

$$\begin{aligned}&\mathbb{E}\,P(\boldsymbol{w}^\star_{\rm reg}) - P(w^o) = \\ &(w^o)^{\mathsf{T}}H^{\mathsf{T}}\Big(I+\frac{1}{\rho N}HH^{\mathsf{T}}\Big)^{-2}Hw^o + \sigma_v^2{\rm Tr}\Big[\Big(H(\rho NI+H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}\Big)^2\Big]\end{aligned}$$

Verify that the expression reduces to the result of Prob. 50.11 as $\rho\to 0$.

51.6 The expression in Prob. 51.5 consists of two terms: the first one depends on $1/\rho$ while the second one varies with ρ. Show that the average excess risk is bounded by

$$\mathbb{E}\,P(\boldsymbol{w}^\star_{\rm reg}) - P(w^o) \le \frac{\rho N}{2}\|w^o\|^2 + \frac{\sigma_v^2}{2\rho N}{\rm Tr}(H^{\mathsf{T}}H)$$

Minimize the bound over ρ and conclude that $\mathbb{E}\,P(\boldsymbol{w}^\star_{\rm reg})-P(w^o) \le \sigma_v\|w^o\|\,\sqrt{{\rm Tr}(H^{\mathsf{T}}H)}$. For which value of ρ is this bound attained?

51.7 Consider the ℓ_1-regularized problem with $\alpha>0$:

$$\underset{w\in\mathbb{R}^M}{\rm argmin}\ \left\{\alpha\|w\|_1\ +\ \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\}$$

Using the vector notation (51.1c), show that the problem is equivalent to solving:

$$\underset{w\in\mathbb{R}^M}{\rm argmin}\ \|w\|_1,\quad \text{subject to } \|d-Hw\|^2\le\epsilon$$

for some $\epsilon\ge 0$. Show that as $\alpha\to 0$, the formulation reduces to the so-called *basis pursuit* problem (which involves an equality constraint):

$$\underset{w\in\mathbb{R}^M}{\rm argmin}\ \|w\|_1,\quad \text{subject to } Hw=d$$

51.8 Establish the validity of expression (51.102) for $\mathbb{S}_\alpha(x)$.

51.9 In this problem, we follow the approach described in the earlier Example 14.10 to express the ℓ_1-regularized least-squares (LASSO) solution in an alternative form. Consider the regularized problem:

$$w^\star_{\rm reg} = \underset{w\in\mathbb{R}^M}{\rm argmin}\ \left\{P_{\rm reg}(w)=\alpha\|w\|_1\ +\ \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n)-y_n^{\mathsf{T}}w\Big)^2\right\}$$

We denote the individual entries of w and y_n by $w={\rm col}\{w_m\}$ and $y_n={\rm col}\{y_{n,m}\}$, respectively, for $m=1,2,\ldots,M$. We also use the notation w_{-m} and $y_{n,-m}$ to refer to the vectors w and y_n with their $m-$th entries excluded.

(a) Verify that, as a function of w_m, the regularized risk can be written as:

$$P_{\rm reg}(w) = a_m\left\{\frac{\alpha}{a_m}|w_m| + \Big(w_m-\frac{c_m}{a_m}\Big)^2\right\} + \text{terms indep. of } w_m$$

where

$$a_m\ \triangleq\ \frac{1}{N}\sum_{n=0}^{N-1}y_{n,m}^2,\qquad c_m\ \triangleq\ \frac{1}{N}\sum_{n=0}^{N-1}y_{n,m}\Big(x(n)-y_{n,-m}^{\mathsf{T}}w_{-m}\Big)$$

(b) Conclude that the minimizer over w_m is given by $\widehat{w}_m = \mathbb{T}_{\alpha/2a_m}(\widehat{c}_m/a_m)$, for $m=1,2,\ldots,M$, and where $\widehat{c}_m$ is given by the same expression as c_m with w_{-m} replaced by $\widehat{w}_{-m}$.

51.10 Replace condition (51.49) by $H^{\mathsf{T}}H = D^2 > 0$, where D is diagonal. Use result (11.35) to show that expression (51.52) is replaced by

$$w^{\star}_{\rm reg} \;=\; \mathrm{sign}(w^{\star}) \odot \Big(|D^2 w^{\star}| - \frac{\alpha N}{2}\Big)_+$$

where the operations $\mathrm{sign}(x)$, $|x|$, and $(a)_+$ are applied elementwise.

51.11 Consider the ℓ_2-regularized risk function $P_{\rm reg}(w) = \rho\|w\|^2 + P(w)$, where $\rho > 0$ and $P(w)$ is some convex risk in w. Show that $P_{\rm reg}(w)$ is strongly convex and, therefore, has a unique global minimum.

51.12 Refer to the equivalent problems (51.94).

(a) Assume $q(w) = \rho\|w\|^2$. Show that τ decreases as ρ increases.

(b) Assume $q(w) = \alpha\|w\|_1 + \rho\|w\|^2$, where $\alpha > 0$ and $\rho > 0$. Show that τ decreases as either α or ρ increases.

51.13 Consider the following ℓ_2-regularized stochastic risk:

$$w^{o}_{\rm reg} = \underset{w\in\mathbb{R}^M}{\mathrm{argmin}} \;\; \Big\{\rho\|w\|_2 \;+\; \mathbb{E}\,(\boldsymbol{x} - \boldsymbol{y}^{\mathsf{T}}w)^2\Big\}$$

Show that $w^{o}_{\rm reg} = R_y(\rho I_M + R_y)^{-1}w^o$, where w^o is the minimizer of the unregularized component, $\mathbb{E}\,(\boldsymbol{x} - \boldsymbol{y}^{\mathsf{T}}w)^2$.

51.14 Consider the following ℓ_1-regularized stochastic risk:

$$w^{o}_{\rm reg} = \underset{w\in\mathbb{R}^M}{\mathrm{argmin}} \;\; \Big\{\; \alpha\|w\|_1 \;+\; \mathbb{E}\,(\boldsymbol{x} - \boldsymbol{y}^{\mathsf{T}}w)^2\Big\}$$

Assume $R_y = \sigma_y^2 I_M$. Show that $w^{o}_{\rm reg} = \mathbb{T}_{\alpha/2\sigma_y^2}\,(w^o)$, where w^o is the minimizer of the unregularized component, $\mathbb{E}\,(\boldsymbol{x} - \boldsymbol{y}^{\mathsf{T}}w)^2$.

51.15 Refer to the ℓ_1-regularized problem (51.38). Verify first that for any scalar $x \in \mathbb{R}$, it holds

$$|x| = \min_{z>0} \;\frac{1}{2}\Big\{\frac{x^2}{z} + z\Big\}$$

Let w_m denote the individual entries of $w \in \mathbb{R}^M$. Conclude that problem (51.38) can be transformed into

$$\min_{w\in\mathbb{R}^M,\{z_m>0\}} \Big\{P_{\rm reg}(w) \;\triangleq\; \alpha\sum_{m=1}^{M}\frac{1}{2}\Big(\frac{w_m^2}{z_m} + z_m\Big) + \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n) - y_n^{\mathsf{T}}w\Big)^2\Big\}$$

Remark. The idea of replacing the regularization factor by smoother forms has been exploited in several works, especially in the context of optimization and image processing — see, for example, Geman Yang (1995), Bach *et al.* (2012), Chan and Liang (2014), and Lanza *et al.* (2015).

51.16 Consider a vector $w \in \mathbb{R}^M$ with individual entries $\{w_m\}$. For any $p \geq 1$ and $\delta \geq 0$, the bridge regression problem in statistics refers to

$$\min_{w\in\mathbb{R}^M} \;\frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n) - y_n^{\mathsf{T}}w\Big)^2, \quad \text{subject to } \sum_{m=1}^{M}|w_m|^p \leq \delta$$

Show that this problem is equivalent to solving

$$\min_{w\in\mathbb{R}^M} \;\Big\{\rho\sum_{m=1}^{M}|w_m|^p \;+\; \frac{1}{N}\sum_{n=0}^{N-1}\Big(x(n) - y_n^{\mathsf{T}}w\Big)^2\Big\}$$

for some $\rho \geq 0$. That is, show that for any $\delta \geq 0$ there exists a $\rho \geq 0$ that makes both problems equivalent to each other (i.e., have the same solution). *Remark.* See the works by Frank and Friedman (1993) and Fu (1998) for a related discussion.

51.17 Refer to the ℓ_1-regularized problem (51.38) and define the quantities $\{d, H\}$ shown in (51.1c), where $H \in \mathbb{R}^{N\times M}$. Let $\{u_m\}$ denote the individual columns of H for $m = 1, 2, \ldots, M$, where each u_m has size $N \times 1$. Let w_m denote the individual entries of $w \in \mathbb{R}^M$. Show that $w^\star$ is a solution of (51.38) if, and only if, for every entry $w_m^\star$ it holds that

$$\begin{cases} |u_m^{\mathsf{T}}(d - Hw^\star)| \leq N\alpha/2, & \text{when } w_m^\star{=}0 \\ \\ u_m^{\mathsf{T}}(d - Hw^\star) = \dfrac{N\alpha}{2}\text{sign}(w_m^\star), & \text{when } w_m^\star \neq 0 \end{cases}$$

Remark. See Bach *et al.* (2012) for a related discussion.

51.18 Derive expressions (51.117a)–(51.117b) for the conjugate functions.

51.19 Assume H is full rank and has dimensions $N \times M$ with $N > M$. Consider the regularized least-squares problem:

$$w^\star \triangleq \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{\rho\|w\|^2 + \|d - Hw\|^2\right\}, \quad \rho > 0$$

and assume $d \notin \mathcal{R}(H)$ and $H^{\mathsf{T}}d \neq 0$. Let $\widetilde{d} = d - Hw^\star$ and introduce the scalar $\eta = \rho\|w^\star\|/\|\widetilde{d}\|$. Verify that $\eta < \|H^{\mathsf{T}}d\|/\|d\|$. *Remark.* See Sayed, Nascimento, and Chandrasekaran (1998) for a related discussion.

51.20 The next two problems are extracted from Sayed (2003,2008). Consider an $N \times M$ full rank matrix H with $N \geq M$, and an $N \times 1$ vector d that does not belong to the column span of H. Let η be a positive real number and consider the set of all matrices δH whose 2−induced norms do not exceed η, $\|\delta H\| \leq \eta$. Now consider the following optimization problem whose solution we denote by $w^\star$:

$$w^\star \triangleq \underset{w\in\mathbb{R}^M}{\text{argmin}} \left\{ \max_{\|\delta H\|\leq\eta} \|d - (H + \delta H)w\| \right\}$$

That is, we seek to minimize the maximum residual over the set $\{\|\delta H\| \leq \eta\}$.

(a) Argue from the conditions of the problem that we must have $N > M$.

(b) Show that the uncertainty set $\{\|\delta H\| \leq \eta\}$ contains a perturbation δH^o such that d is orthogonal to $(H + \delta H^o)$ if, and only if, $\eta \geq \|H^{\mathsf{T}}d\|/\|d\|$.

(c) Show that the above optimization problem has a unique solution at $w^\star = 0$ if, and only if, the condition on η in part (b) holds.

Remark. For more details on such robust formulations, see Chandrasekaran *et al.* (1997,1998), Sayed, Nascimento, and Chandrasekaran (1998), and Sayed, Nascimento, and Cipparrone (2002).

51.21 Consider an $N \times M$ full rank matrix H with $N \geq M$, and an $N \times 1$ vector d that does not belong to the column span of H.

(a) For any nonzero $M \times 1$ column vector w, show that the following rank-one modification of H continues to have full rank for any positive real number η:

$$H(w) \triangleq H - \eta\frac{d - Hw}{\|d - Hw\|}\frac{w^{\mathsf{T}}}{\|w\|}$$

(b) Verify that $\|d - H(w)w\| = \|y - Hw\| + \eta\|w\|$, and that the vectors $d - H(w)w$ and $d - Hw$ are collinear and point in the same direction (that is, one is a positive multiple of the other).

(c) Show that $\|d - H(w)w\| = \max_{\|\delta H\|\leq\eta} \|d - (H + \delta H)w\|$.

(d) Show that the optimization problem

$$\min_{w\in\mathbb{R}^M} \max_{\|\delta H\|\leq\eta} \|d - (H + \delta H)w\|$$

has a nonzero solution $w^\star$ if, and only if, $\eta < \|H^{\mathsf{T}}d\|/\|d\|$.

(e) Show that $w^\star$ is a nonzero solution of the optimization problem in part (d) if, and only, if $H^{\mathsf{T}}(w^\star)(d - Hw^\star) = 0$. That is, the residual vector $d - Hw^\star$ should be orthogonal to the perturbed matrix $H(w^\star)$. Show further that this condition is equivalent to $H^{\mathsf{T}}(w^\star)(d - H(w^\star)w^\star) = 0$.

(f) Assume two nonzero solutions $w_1^\star$ and $w_2^\star$ exist that satisfy the orthogonality condition of part (e). Argue that $H^{\mathsf{T}}(w_2^\star)(d - H(w_2^\star)w_1^\star) = 0$, and conclude that $w_1^\star = w_2^\star$ so that the solution is unique.

Remark. For further details, see Sayed, Nascimento, and Chandrasekaran (1998).

51.A CONSTRAINED FORMULATIONS FOR REGULARIZATION

In this appendix, we first establish the equivalence between problems (51.16) and (51.26) for ℓ_2-regularized least-squares, and between problems (51.38)–(51.39) and (51.44) for ℓ_1 and elastic-net regularized least-squares. Then we extend the conclusion to other regularized convex risks, besides least-squares. Although it would have been sufficient to treat the general case right away, we prefer to explain the equivalence in a gradual manner for the benefit of the reader, starting with quadratic risks. To establish the equivalence, we will appeal to the Lagrange and KKT multiplier arguments from Sec. 9.1.

51.A.1 Quadratic Risks

We start with the ℓ_2-regularized least-squares risk.

ℓ_2-regularization

To begin with, we identify the smallest value for τ. We already know that the solution to the ℓ_2-regularized problem (51.16) is given by

$$w_{\text{reg}}^\star \;=\; (\rho N I_M + H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}d \tag{51.82}$$

Now, consider the constrained problem (51.26) for some $\tau > 0$. The unregularized risk $P(w)$ is quadratic in w and is therefore convex and continuously differentiable. The constraint $\rho\|w\|^2 \leq \tau$ defines a convex set in $\mathbb{R}^M$. We are therefore faced with the problem of minimizing a convex function over a convex domain. It is straightforward to verify that problems of this type can only have global minima — see the argument after (9.10). For the solution $w_{\text{reg}}^\star$ defined by (51.82) to be included in the search domain $\rho\|w\|^2 \leq \tau$, it is necessary for the value of τ to satisfy $\tau \geq \rho\|w_{\text{reg}}^\star\|^2$. This argument shows that the smallest value for τ is

$$\tau \;=\; \rho\|w_{\text{reg}}^\star\|^2 \;=\; \rho\|(\rho N I_M + H^{\mathsf{T}}H)^{-1}H^{\mathsf{T}}d\|^2 \tag{51.83}$$

Actually, the regularized solution, $w_{\text{reg}}^\star$, will lie on the boundary of the set $\|w\|^2 \leq \tau/\rho$. At the same time, the constraint set will *exclude* any of the solutions, $w^\star$, to the original unregularized solution from (51.1b). This is because $\|w^\star\| > \|w_{\text{reg}}^\star\|$, as already revealed by (51.24).

Let $w_{\text{cons}}^\star$ denote a solution to the constrained problem (51.26) for the above value of τ. We want to verify that this solution agrees with $w_{\text{reg}}^\star$. We appeal to the KKT conditions from Sec. 9.1. Note first that problem (51.26) does not involve any equality constraints and has only one inequality constraint of the form

$$g(w) \;\triangleq\; \rho\|w\|^2 - \tau \;\leq\; 0 \tag{51.84}$$

We introduce the Lagrangian function

$$\mathcal{L}(w,\lambda) \;=\; P(w) + \lambda(\rho\|w\|^2 - \tau), \;\; \lambda \geq 0 \tag{51.85}$$

and let $w^\star_{\rm reg}(\lambda)$ denote a minimizer for it. Strong duality holds because Slater condition (9.58a) is satisfied, i.e., there exists a $\bar{w}$ such that $g(\bar{w}) < 0$ (e.g., $\bar{w} = 0$). The KKT conditions (9.28a)–(9.28e) then state that $w^\star_{\rm reg}(\lambda)$ agrees with $w^\star_{\rm cons}$ if, and only if, the following conditions hold for some scalar λ and $w^\star_{\rm reg}(\lambda)$:

$$\rho\|w^\star_{\rm reg}(\lambda)\|^2 - \tau \leq 0, \qquad \text{(feasibility of primal problem)} \tag{51.86a}$$

$$\lambda \geq 0, \qquad \text{(feasibility of dual problem)} \tag{51.86b}$$

$$\lambda\left(\rho\|w^\star_{\rm reg}(\lambda)\|^2 - \tau\right) = 0, \qquad \text{(complementary condition)} \tag{51.86c}$$

$$\nabla_w \Big\{\lambda(\rho\|w\|^2 - \tau) \;+\; P(w)\Big\}_{w=w^\star_{\rm reg}(\lambda)} = 0 \tag{51.86d}$$

If we select $\lambda = 1$, then $w^\star_{\rm reg}(\lambda) = w^\star_{\rm reg}$ and the KKT conditions are satisfied at these values for $\tau = \rho\|w^\star_{\rm reg}\|^2$. It follows that $w^\star_{\rm cons} = w^\star_{\rm reg}$.

ℓ_1 and elastic-net regularization

We can extend the argument for other regularization factors, such as $q(w) = \alpha\|w\|_1$ or $q(w) = \alpha\|w\|_1 + \rho\|w\|^2$. Let $w^\star_{\rm reg}$ denote the minimizer for either regularized risk (51.38) or (51.39); the argument applies to both cases. It follows that the smallest value for τ should be:

$$\tau = q(w^\star_{\rm reg}) \tag{51.87}$$

Let $w^\star_{\rm cons}$ denote a solution to the constrained problem (51.44) for the above value of τ. We want to verify that this solution agrees with $w^\star_{\rm reg}$. We again appeal to the KKT conditions from Sec. 9.1. Note first that either problem (51.38) or (51.39) does not involve any equality constraints and has only one inequality constraint of the form

$$g(w) \;\stackrel{\Delta}{=}\; q(w) - \tau \;\leq\; 0 \tag{51.88}$$

We introduce the Lagrangian function

$$\mathcal{L}(w,\lambda) \;=\; P(w) + \lambda(q(w) - \tau), \;\; \lambda \geq 0 \tag{51.89}$$

and let $w^\star_{\rm reg}(\lambda)$ denote a minimizer for it. Strong duality holds because Slater condition (9.58a) is satisfied, i.e., there exists a $\bar{w}$ such that $g(\bar{w}) < 0$ (e.g., $\bar{w} = 0$). The KKT conditions (9.28a)–(9.28e) then state that $w^\star_{\rm reg}(\lambda)$ agrees with $w^\star_{\rm cons}$ if, and only if, the following conditions hold for some scalar λ and $w^\star_{\rm reg}(\lambda)$:

$$q(w^\star_{\rm reg}(\lambda)) - \tau \leq 0, \qquad \text{(feasibility of primal problem)} \tag{51.90a}$$

$$\lambda \geq 0, \qquad \text{(feasibility of dual problem)} \tag{51.90b}$$

$$\lambda\left(q(w^\star_{\rm reg}(\lambda)) - \tau\right) = 0, \qquad \text{(complementary condition)} \tag{51.90c}$$

$$0 \in \partial\Big\{q(w) \;+\; P(w)\Big\}_{w=w^\star_{\rm reg}(\lambda)} \tag{51.90d}$$

If we select $\lambda = 1$, then $w^\star_{\rm reg}(\lambda) = w^\star_{\rm reg}$ and the KKT conditions are satisfied at these values for $\tau = q(w^\star_{\rm reg})$. It follows that $w^\star_{\rm cons} = w^\star_{\rm reg}$.

51.A.2 Other Convex Risks

The discussion in the body of the chapter reveals that regularization has several benefits: it resolves ambiguities by ensuring unique solutions and counters ill-conditioning and overfitting. Naturally, these favorable conditions come at the expense of introducing bias: the achievable minimum risk (or training error) is higher under regularization

than it would be in the absence of regularization. These various properties have been established so far for the case of least-squares risks. We argue now that regularization ensures similar properties for other convex risk functions besides quadratic risks. Thus, more generally, we let $P(w)$ denote any convex risk function, differentiable or not, and introduce its regularized version:

$$P_{\rm reg}(w) \triangleq q(w) + P(w) \tag{51.91}$$

where the penalty term, $q(w)$, is also assumed to be convex in w such as the choices introduced earlier in (51.15). In the chapter, we considered one choice for $P(w)$, namely, the quadratic risk (51.92a). Later, when we study learning algorithms, other convex empirical risks will arise (such as logistic risks, exponential risks, hinge risks and others), in which case the results of the current appendix will be applicable; some of the risks will also be non-differentiable. Examples include empirical risks of the form:

$$P(w) = \frac{1}{N}\sum_{n=0}^{N-1}\left(x(n) - y_n^{\mathsf{T}}w\right)^2, \qquad \text{(quadratic risk)} \tag{51.92a}$$

$$P(w) = \frac{1}{N}\sum_{n=0}^{N-1}\ln\left(1 + e^{-x(n)y_n^{\mathsf{T}}w}\right), \qquad \text{(logistic risk)} \tag{51.92b}$$

$$P(w) = \frac{1}{N}\sum_{n=0}^{N-1}\max\left\{0,\, -x(n)y_n^{\mathsf{T}}w\right\}, \qquad \text{(Perceptron risk)} \tag{51.92c}$$

$$P(w) = \frac{1}{N}\sum_{n=0}^{N-1}\max\left\{0,\, 1 - x(n)y_n^{\mathsf{T}}w\right\}, \qquad \text{(hinge risk)} \tag{51.92d}$$

Uniqueness of solution. We focus on general optimization problems of the form (51.91), where $P(w)$ is convex in w (but need not be differentiable) and $q(w)$ is one of the convex penalty terms considered before in (51.15).

The first property to note is that whenever $P(w)$ is convex in w, the ℓ_2-regularized version (i.e., when $q(w) = \rho\|w\|^2$), will be strongly convex for any $\rho > 0$ and, therefore, $P_{\rm reg}(w)$ will have a unique global minimum, $w^\star_{\rm reg}$. The strong convexity of $P_{\rm reg}(w)$ in this case follows from the fact that $\rho\|w\|^2$ is itself strongly-convex — see Prob. 51.11. Therefore, ridge regression ensures a unique global minimizer; a similar conclusion can be established when elastic-net regularization is applied for any convex empirical risk $P(w)$. For ℓ_1-regularization, a unique global minimizer will be guaranteed when $P(w)$ happens to be strictly or strongly convex; in this case, convexity of $P(w)$ alone is not sufficient because the penalty $\alpha\|w\|_1$ is convex but not strictly convex. In the sequel, we *assume that the regularized risk* $P_{\rm reg}(w)$ *has a unique global minimizer.*

Promoting smaller solutions. Let $w^\star$ denote a global minimizer for the unregularized convex risk, $P(w)$. This minimizer need not be unique since $P(w)$ is only assumed to be convex but not necessarily strongly convex. Let $w^\star_{\rm reg}$ denote the global minimizer for the regularized risk, $P_{\rm reg}(w)$. This minimizer is unique. Now, since $w^\star_{\rm reg}$ minimizes $P_{\rm reg}(w)$, we have

$$\begin{aligned}
&q(w^\star_{\rm reg}) + P(w^\star_{\rm reg}) \le q(w^\star) + P(w^\star)\\
&\Longrightarrow q(w^\star_{\rm reg}) - q(w^\star) \le P(w^\star) - P(w^\star_{\rm reg})\\
&\overset{(a)}{\Longrightarrow} q(w^\star_{\rm reg}) - q(w^\star) \le 0\\
&\Longleftrightarrow q(w^\star_{\rm reg}) \le q(w^\star)
\end{aligned} \tag{51.93}$$

where step (a) is because $w^\star$ minimizes the unregularized risk, $P(w)$.

Constrained formulation. We assume the regularized risk has a unique global minimizer.

We also assume that the *Slater condition (9.58a) holds*, i.e., there exists a $\bar{w}$ such that $g(\bar{w}) < 0$, which is equivalent to $q(\bar{w}) < \tau$ where $\tau = q(w^{\star}_{\rm reg})$. This can be satisfied, for example, at $\bar{w} = 0$ for the penalty terms considered before in (51.15). Then, the same KKT argument used earlier in this appendix under ℓ_1 and elastic-net regularization shows that the following two problems are equivalent (meaning they have the same solution vectors):

$$\begin{cases} w^{\star}_{\rm reg} = \underset{w\in\mathbb{R}^M}{\rm argmin}\ \Big\{q(w) + P(w)\Big\} \iff \\ w^{\star}_{\rm cons} = \underset{w\in\mathbb{R}^M}{\rm argmin}\ \ P(w)\ \ \text{subject to } q(w) \le q(w^{\star}_{\rm reg}) \end{cases} \tag{51.94}$$

51.B EXPRESSION FOR LASSO SOLUTION

In this appendix, we establish Theorem 51.1 for the solution of the LASSO problem under ℓ_1 and elastic-net regularization using a duality argument patterned after the derivation in Chen, Towfic, and Sayed (2015); other related arguments appear in Mota *et al.* (2012,2013).

We assume first that $\rho \neq 0$. We start by introducing the Lagrangian function:

$$\mathcal{L}(w, z, \lambda) \triangleq \frac{1}{N}\|d - z\|^2 + q(w) + \lambda^{\mathsf{T}}(z - Hw) \tag{51.95}$$

where $\lambda \in \mathbb{R}^N$ is the dual variable (or Lagrange multiplier). The dual function is defined by minimizing $\mathcal{L}(w, z, \lambda)$ over $\{w, z\}$:

$$\begin{aligned} \mathcal{D}(\lambda) &\triangleq \min_{w,z}\ \mathcal{L}(w, z, \lambda), \qquad (\textbf{dual function}) \\ &= \min_{z}\left\{\frac{1}{N}\|d - z\|^2 + \lambda^{\mathsf{T}} z\right\} + \min_{w}\left\{q(w) - \lambda^{\mathsf{T}} H w\right\} \end{aligned} \tag{51.96}$$

where we are grouping separately the terms that depend on z and w. Once this dual function is determined, as shown by future expression (51.107), maximizing it leads to the optimal value for λ — see Eq. (51.62):

$$\lambda^o = \underset{\lambda\in\mathbb{R}^N}{\rm argmax}\ \mathcal{D}(\lambda) \tag{51.97}$$

Strong duality will then imply that we can determine the optimal solutions for $\{w, z\}$ for problem (51.60) by using this value for λ^o, namely, by solving:

$$z^o \triangleq \underset{z\in\mathbb{R}^N}{\rm argmin}\left\{\frac{1}{N}\|d - z\|^2 + (\lambda^o)^{\mathsf{T}} z\right\} \implies z^o = d - \frac{N}{2}\lambda^o \tag{51.98a}$$

$$w^{\star}_{\rm reg} \triangleq \underset{w\in\mathbb{R}^M}{\rm argmin}\ \Big\{q(w) - (\lambda^o)^{\mathsf{T}} H w\Big\} \tag{51.98b}$$

Expression (51.98a) shows how z^o is determined from λ^o. We still need to show how to solve (51.98b) and determine the regularized solution in terms of λ^o. We can pursue

this task by appealing to result (51.47). Indeed, note that

$$\begin{aligned} w^{\star}_{\rm reg} &= \operatorname*{argmin}_{w\in\mathbb{R}^M} \Big\{ q(w) - (\lambda^o)^{\mathsf{T}} H w \Big\} \\ &= \operatorname*{argmin}_{w\in\mathbb{R}^M} \Big\{ \alpha\|w\|_1 + \rho\|w\|^2 - (\lambda^o)^{\mathsf{T}} H w \Big\} \\ &= \operatorname*{argmin}_{w\in\mathbb{R}^M} \Big\{ \frac{\alpha}{\rho}\|w\|_1 + \|w\|^2 - \frac{1}{\rho}(\lambda^o)^{\mathsf{T}} H w \Big\} \\ &= \operatorname*{argmin}_{w\in\mathbb{R}^M} \left\{ \frac{\alpha}{\rho}\|w\|_1 + \left\| w - \frac{1}{2\rho} H^{\mathsf{T}}\lambda^o \right\|^2 - \frac{1}{(2\rho)^2}\left\| H^{\mathsf{T}}\lambda^o \right\|^2 \right\} \end{aligned} \tag{51.99}$$

Using result (51.47) we conclude that (51.61) holds.

Determining the dual variable $\boldsymbol{\lambda^o}$. To complete the argument, we still need to determine λ^o, which is the maximizer for the dual function $\mathcal{D}(\lambda)$ defined by (51.96). We first determine $\mathcal{D}(\lambda)$. From (51.96) we observe that we need to minimize two separate terms, one over z and one over w.

We already know from the above argument that for any λ:

$$\operatorname*{argmin}_{w\in\mathbb{R}^M} \Big\{ q(w) - \lambda^{\mathsf{T}} H w \Big\} \implies w^{\star}_{\lambda} = \frac{1}{2\rho}\,\mathbb{T}_{\alpha}(H^{\mathsf{T}}\lambda) \tag{51.100}$$

where we are denoting the minimizer for a generic λ by the notation $w^{\star}_{\lambda}$. Consequently, the minimum value for this first minimization is given by

$$\begin{aligned} q(w^{\star}_{\lambda}) - \lambda^{\mathsf{T}} H w^{\star}_{\lambda} &= \alpha\|w^{\star}_{\lambda}\|_1 + \rho\|w^{\star}_{\lambda}\|^2 - \lambda^{\mathsf{T}} H w^{\star}_{\lambda} \\ &\overset{(51.100)}{=} \frac{1}{2\rho}\left(\alpha\|\mathbb{T}_{\alpha}(H^{\mathsf{T}}\lambda)\|_1 + \frac{1}{2}\|\mathbb{T}_{\alpha}(H^{\mathsf{T}}\lambda)\|^2 - \lambda^{\mathsf{T}} H \mathbb{T}_{\alpha}(H^{\mathsf{T}}\lambda) \right) \end{aligned} \tag{51.101}$$

To simplify the notation, we let for any vector x:

$$\mathbb{S}_{\alpha}(x) \triangleq -\alpha\|\mathbb{T}_{\alpha}(x)\|_1 - \frac{1}{2}\|\mathbb{T}_{\alpha}(x)\|^2 + x^{\mathsf{T}}\mathbb{T}_{\alpha}(x) \tag{51.102}$$

Then, it is verified in Prob. 51.8 that

$$\mathbb{S}_{\alpha}(x) = \frac{1}{2}\|\mathbb{T}_{\alpha}(x)\|^2 \tag{51.103}$$

In this way, we can rewrite the minimum value (51.101) more compactly as

$$q(w^{\star}_{\lambda}) - \lambda^{\mathsf{T}} H w^{\star}_{\lambda} = -\frac{1}{2\rho}\mathbb{S}_{\alpha}(H^{\mathsf{T}}\lambda) = -\frac{1}{4\rho}\|\mathbb{T}_{\alpha}(H^{\mathsf{T}}\lambda)\|^2 \tag{51.104}$$

For illustration purposes, Figure 51.6 plots the soft-thresholding functions $\mathbb{T}_{\alpha}(x)$ and $\mathbb{S}_{\alpha}(x)$ for $\alpha = 1$ and a scalar argument, x.

Let us now consider the first minimization in (51.96) for any λ:

$$\min_{z\in\mathbb{R}^N} \left\{ \frac{1}{N}\|d - z\|^2 + \lambda^{\mathsf{T}} z \right\} \implies \widehat{z}_{\lambda} = d - \frac{N}{2}\lambda \tag{51.105}$$

where we are denoting the minimizer for a generic λ by the notation $\widehat{z}_{\lambda}$. Consequently, the minimum value for this minimization is given by

$$\frac{1}{N}\|d - \widehat{z}_{\lambda}\|^2 + \lambda^{\mathsf{T}}\widehat{z}_{\lambda} = \lambda^{\mathsf{T}} d - \frac{N}{4}\|\lambda\|^2 \tag{51.106}$$

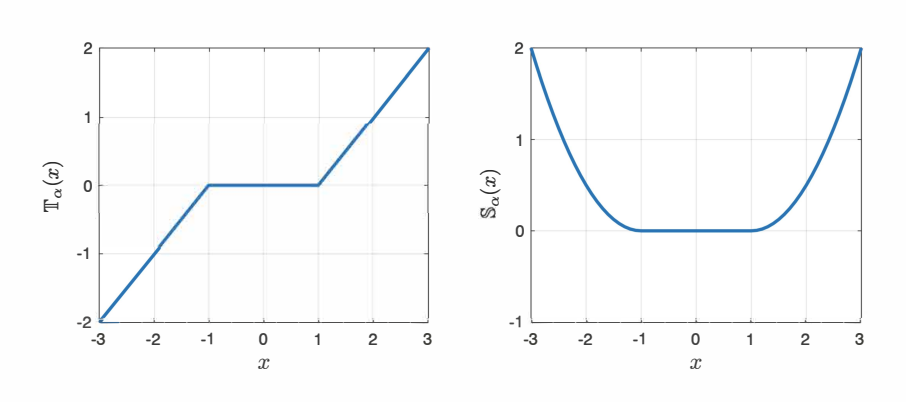

Figure 51.6 Plots of the soft-thresholding functions $\mathbb{T}_\alpha(x)$ and $\mathbb{S}_\alpha(x)$ for $\alpha = 1$.

Adding this result to (51.104) we find that the dual function is given by

$$\mathcal{D}(\lambda) \;=\; \lambda^{\mathsf{T}} d - \frac{N}{4}\|\lambda\|^2 - \frac{1}{4\rho}\left\|\mathbb{T}_\alpha\left(H^{\mathsf{T}}\lambda\right)\right\|^2 \tag{51.107}$$

It can be verified that this function is strongly-concave and, therefore, has a unique maximum (see next example). The desired dual variable, λ^o, is therefore given by (51.62).

The proof technique used so far requires $\rho > 0$. This condition was used to complete the squares in step (51.99). We now explain how to handle the situation $\rho = 0$, which corresponds to pure ℓ_1-regularization. The solution will continue to be given by expression (51.61), except that λ^o will be found by solving the projection problem (51.112). The details are as follows.

We revisit step (51.99) when $\rho = 0$ and note that it reduces to solving a problem of the form:

$$\min_{w\in\mathbb{R}^M}\left\{\alpha\|w\|_1 - \lambda^{\mathsf{T}} H w\right\} \tag{51.108}$$

Let $\mathcal{C}$ denote the convex set of vectors satisfying $\|x\|_\infty \leq 1$. We established earlier in Table 8.4 and Prob. 8.55 the following conjugate pair:

$$r(w) = \|w\|_1 \;\;\Longrightarrow\;\; r^\star(x) = \mathbb{I}_{C,\infty}[x] \tag{51.109}$$

where the notation $\mathbb{I}_{C,\infty}[x]$ represents the indicator function relative to set $\mathcal{C}$: it assumes the value zero if $x \in \mathcal{C}$ and $+\infty$ otherwise. In light of definition (51.114) for the conjugate function, we find that the minimum value of problem (51.108) is given by

$$\min_{w\in\mathbb{R}^M}\left\{\alpha\|w\|_1 - \lambda^{\mathsf{T}} H w\right\} \;=\; -\mathbb{I}_{C,\infty}[H^{\mathsf{T}}\lambda/\alpha] \tag{51.110}$$

Adding this value to (51.106) we find that the dual function is now given by

$$\mathcal{D}(\lambda) \;=\; \lambda^{\mathsf{T}} d - \frac{N}{4}\|\lambda\|^2 - \mathbb{I}_{C,\infty}[H^{\mathsf{T}}\lambda/\alpha] \tag{51.111}$$

Maximizing $\mathcal{D}(\lambda)$ over λ results in λ^o. To do so, we complete the squares over λ to find that the maximization of $\mathcal{D}(\lambda)$ is equivalent to solving:

$$\lambda^o = \operatorname*{argmin}_{\lambda\in\mathbb{R}^N}\left\|\lambda - \frac{2}{N}d\right\|^2, \quad \text{subject to } \|H^{\mathsf{T}}\lambda\|_\infty \leq \alpha \tag{51.112}$$

The minimizer λ^o is obtained by projecting $\frac{2}{N}d$ onto the set of all vectors λ satisfying $\|H^{\mathsf{T}}\lambda\|_\infty \leq \alpha$. Using $z = Hw$ and $z^o = d - \frac{N}{2}\lambda^o$, we conclude that the optimal solution $w^\star_{\rm reg}$ also satisfies the equation

$$Hw^\star_{\rm reg} = d - \frac{N}{2}\lambda^o \;=\; \frac{2}{N}\underbrace{\left(\frac{2}{N}d - \lambda^o\right)}_{\text{residual}} \tag{51.113}$$

in terms of the residual resulting from projecting $\frac{2}{N}d$ onto the set $\|H^{\mathsf{T}}\lambda\|_\infty \leq \alpha$.

Example 51.6 (**Duality and conjugate functions**) There is an alternative way to arrive at the same expression (51.62) by calling upon the concept of *conjugate* functions, also called Fenchel conjugate functions. This alternative argument is useful for situations (other than least-squares) when explicit expressions for the individual minimum values (51.104) and (51.106) may not be directly available but can be expressed in terms of conjugate functions.

We first recall the definition of conjugate functions from (8.83). Consider an arbitrary function $r(w) : \mathbb{R}^M \to \mathbb{R}$ with domain $\mathrm{dom}(r)$; the function $r(w)$ need not be convex. Its conjugate function is denoted by $r^\star(\lambda) : \mathbb{R}^M \to \mathbb{R}$ and is defined as:

$$r^\star(\lambda) \;\triangleq\; \sup_{w\in\mathbb{R}^M} \Big\{\lambda^{\mathsf{T}}w - r(w)\Big\}, \quad \lambda \in \mathcal{Y} \tag{51.114}$$

where $\mathcal{Y}$ denotes the set of all λ where the supremum operation is finite. It can be verified that $r^\star(\lambda)$ is convex regardless of whether $r(w)$ is convex or not. Likewise, the set $\mathcal{Y}$ is a convex set — recall Prob. 8.47 and Table 8.4. If $r(w)$ happens to be strongly-convex, then $\mathcal{Y} = \mathbb{R}^M$ (i.e., the sup is finite for all λ).

Now, consider the quadratic function $f(w) = \|w\|^2$ and observe that the dual function $\mathcal{D}(\lambda)$ in (51.96) can be written as:

$$\begin{aligned}
\mathcal{D}(\lambda) &= -\sup_{z\in\mathbb{R}^N}\left(-\lambda^{\mathsf{T}}z - \frac{1}{N}\|d-z\|^2\right) \;-\; \sup_{w\in\mathbb{R}^M}\Big(\lambda^{\mathsf{T}}Hw - q(w)\Big) \\
&= -\sup_{z\in\mathbb{R}^N}\left(\lambda^{\mathsf{T}}(d-z) - \frac{1}{N}\|d-z\|^2 - \lambda^{\mathsf{T}}d\right) \;-\; q^\star(H^{\mathsf{T}}\lambda) \\
&= -\sup_{s\in\mathbb{R}^N}\left(\lambda^{\mathsf{T}}s - \frac{1}{N}\|s\|^2 - \lambda^{\mathsf{T}}d\right) \;-\; q^\star(H^{\mathsf{T}}\lambda), \;\; s \;\triangleq\; d-z \\
&= -\sup_{s\in\mathbb{R}^N}\left(\lambda^{\mathsf{T}}s - \frac{1}{N}\|s\|^2\right) \;+\; \lambda^{\mathsf{T}}d \;-\; q^\star(H^{\mathsf{T}}\lambda) \\
&= -\frac{1}{N}\sup_{s\in\mathbb{R}^N}\left(N\lambda^{\mathsf{T}}s - \|s\|^2\right) \;+\; \lambda^{\mathsf{T}}d \;-\; q^\star(H^{\mathsf{T}}\lambda) \\
&= -\frac{1}{N}f^\star(N\lambda) + \lambda^{\mathsf{T}}d \;-\; q^\star(H^{\mathsf{T}}\lambda)
\end{aligned} \tag{51.115}$$

where $f^\star(\lambda)$ and $q^\star(\lambda)$ denote the conjugate functions of

$$f(w) = \|w\|^2, \quad q(w) = \alpha\|w\|_1 + \rho\|w\|^2 \tag{51.116}$$

Both functions, $f(w)$ and $q(w)$, are strongly-convex and, therefore, the domains of their conjugate functions are the entire space, $\mathbb{R}^M$. Moreover, since $f(w)$ and $q(w)$ are strongly-convex and differentiable, it follows from the properties of conjugate functions that $f^\star(\lambda)$ and $q^\star(\lambda)$ are themselves strongly-convex and differentiable — recall Table 8.4. This implies that $\mathcal{D}(\lambda)$ is strongly-concave (i.e., its negative is strongly-convex),

differentiable, and has a unique maximizer, λ^o.

It can be verified that the conjugate functions for $f(w)$ and $q(w)$ are given by — see Prob. 51.18:

$$f(w) = \|w\|^2 \implies f^\star(\lambda) = \frac{1}{4}\|\lambda\|^2 \tag{51.117a}$$

$$q(w) = \alpha\|w\|_1 + \rho\|w\|^2 \implies q^\star(\lambda) = \frac{1}{4\rho}\|\mathbb{T}_\alpha(\lambda)\|^2 \tag{51.117b}$$

Substituting into (51.115), we find that the dual function is given by (51.107).

REFERENCES

Bach, F., R. Jenatton, J. Mairal, and G. Obozinski (2012), "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106.

Bjorck, A. (1996), *Numerical Methods for Least Squares Problems,* SIAM, PA.

Chan, R. H. and H. X. Liang (2014), "Half-quadratic algorithm for $\ell_p - \ell_q$ problems with applications to TV-1 image restoration and compressive sensing," pp. 78–103, *Lecture Notes in Comput. Sci.* 8293, Springer, Berlin.

Chandrasekaran, S., G. Golub, M. Gu, and A. H. Sayed (1997), "Parameter estimation in the presence of bounded modeling errors," *IEEE Signal Processing Letters*, vol. 4, no. 7, pp. 195–197.

Chandrasekaran, S., G. Golub, M. Gu, and A. H. Sayed (1998), "Parameter estimation in the presence of bounded data uncertainties," *SIAM. J. Matrix Anal. Appl.,* vol. 19, no. 1, pp. 235–252.

Chen, S. and D. Donoho (1994), "Basis pursuit," *Proc. Asilomar Conference on Signals, Systems and Computers*, pp. 41–44, Pacific Grove, CA.

Chen, S. S., D. L. Donoho, and M. A. Saunders (1998), "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1 pp. 33–61. Republished in *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001. An earlier draft has been available earlier since 1995 as a technical report, Department of Statistics, Stanford University.

Chen, S., D. Donoho, and M. Saunders (2001), "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159.

Chen, J., Z. J. Towfic, and A. H. Sayed (2015), "Dictionary learning over distributed models," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 1001–1016.

Claerbout, J. F. and F. Muir (1973), "Robust modeling with erratic data,' *Geophysics,* vol. 38, no. 5, pp. 826–844.

Dahleh, M. A. and I. Diaz-Bobillo (1995), *Control of Uncertain Systems: A Linear Programming Approach,* Prentice Hall, NY.

Dahleh, M. A. and J. B. Pearson (1986), "ℓ_1-optimal feedback controllers for discrete-time systems," *Proc. Amer. Control Conf.* (ACC), pp. 1964–1968, Seattle, WA.

Dahleh, M. A. and J. B. Pearson (1987), "ℓ_1-optimal feedback controllers for MIMO discrete-time systems," *IEEE Trans. Automat. Contr.*, vol. 32, pp. 314–322.

De Mol, C., E. De Vito, and L. Rosasco (2009), "Elastic-net regularization in learning theory," *J. Complexity*, vol. 25, no. 2, pp. 201–230.

El Ghaoui, L. and H. Lebret (1997), "Robust solutions to least-squares problems with uncertain data," *SIAM. J. Matrix Anal. Appl.,* vol. 18, no. 4, pp. 1035–1064.

Engl, H. W., M. Hanke, and A. Neubauer (1996), *Regularization of Inverse Problems*, Kluwer.

Frank, I. E. and J. H. Friedman (1993), "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, pp. 109–148.

Fu, W. J. (1998), "Penalized regressions: The bridge versus the Lasso," *J. Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397–416.

Geman, D. and C. Yang (1995), "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Processing*, vol. 4, no. 7, pp. 932–946.

Golub, G. H. and C. F. Van Loan (1996), *Matrix Computations*, 3rd edition, The John Hopkins University Press, MD.

Hansen, P. C. (1997), *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, PA.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning*, 2nd edition, Springer, NY.

Hoerl, A. E. (1959), "Optimum solution of many variables equations," *Chemical Engineering Progress*, vol. 55, pp. 69–78.

Hoerl, A. E. (1962), "Application of ridge analysis to regression problems," *Chemical Engineering Progress*, vol. 58, pp. 54–59.

Hoerl, A. E. and R. W. Kennard (1970), "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67.

Hoerl, R. W. (1985), "Ridge analysis 25 years later," *The American Statistician*, vol. 39, pp. 186–192.

Lanza, A., S. Morigi, L. Reichel, and F. Sgallari (2015), "A generalized Krylov subspace method for $\ell_p - \ell_q$ minimization," *SIAM J. Sci. Computing*, vol. 37, no. 5, pp. 30–50.

Levy, S. and P. K. Fullagar (1981), "Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution," *Geophysics*, vol. 46, no. 9, pp. 1235–1243.

Mota, J., J. Xavier, P. Aguiar, and M. Puschel (2012), "Distributed basis pursuit," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1942–1956, Apr. 2012.

Mota, J., J. Xavier, P. Aguiar, and M. Puschel (2013), "D-ADMM: A communication-efficient distributed algorithm for separable optimization," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2718–2723.

Neumaier, A. (1998), "Solving ill-conditioned and singular linear systems: A tutorial on regularization," *SIAM Review*, vol. 40, no. 3, pp. 636–666.

Oldenburg, D., W. T. Scheuer, and S. Levy (1983), "Recovery of the acoustic impedance from reflection seismograms", *Geophysics*, vol. 48, no. 10, pp. 1318–1337.

Phillips, D. L. (1962), "A technique for the numerical solution of certain integral equations of the first kind," *Journal of the ACM*, vol. 9, no. 1, pp. 84–97.

Santosa, F. and W. W. Symes (1986), "Linear inversion of band-limited reflection seismograms," *SIAM J. Sci. Statist. Comput.*, vol. 7, no. 4, pp. 1307–1330.

Sayed, A. H. and H. Chen (2002), "A uniqueness result concerning a robust regularized least-squares solution," *Systems and Control Letters*, vol. 46, pp. 361–369.

Sayed, A. H., V. Nascimento, and F. A. M. Cipparrone (2002), "A regularized robust design criterion for uncertain data," *SIAM J. Matrix Anal. Appl.*, vol. 23, no. 4, pp. 1120–1142.

Sayed, A. H., V. H. Nascimento, and S. Chandrasekaran (1998), "Estimation and control with bounded data uncertainties," *Linear Algebra and Its Applications*, vol. 284, pp. 259–306.

Taylor, H. L., S. C. Banks, and J. F. McCoy (1979), "Deconvolution with the ℓ_1 norm," *Geophysics,* vol. 44, no. 1, pp. 39–52.

Tibshirani, R. (1996b), "Regression shrinkage and selection via the Lasso," *J. Royal Stat. Society*, Series B, vol. 58, no. 1, pp. 267–288.

Tikhonov, A. N. (1963), "Solution of incorrectly formulated problems and the regularization method," *Soviet Math. Dokl.*, vol. 4, pp. 1035–1038.

Tikhonov, A. N. and V. Y. Arsenin (1977), *Solutions of Ill-Posed Problems*, Winston, NY.

Vidyasagar, M. (1986), "Optimal rejection of persistent bounded disturbances," *IEEE Trans. Aut. Control*, vol. 31, no. 6, pp. 527–534.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, PA.

Zou, H. and T. Hastie (2005), "Regularization and variable selection via the elastic net," *J. Royal Stat. Society*, Series B, pp. 301–320.