9

Decomposability and restricted strong convexity

In Chapter 7, we studied the class of sparse linear models, and the associated use of ℓ_1-regularization. The basis pursuit and Lasso programs are special cases of a more general family of estimators, based on combining a cost function with a regularizer. Minimizing such an objective function yields an estimation method known as an *M-estimator*. The goal of this chapter is to study this more general family of regularized M-estimators, and to develop techniques for bounding the associated estimation error for high-dimensional problems. Two properties are essential to obtaining consistent estimators in high dimensions: decomposability of the regularizer, and a certain type of lower restricted curvature condition on the cost function.

9.1 A general regularized M-estimator

Our starting point is an indexed family of probability distributions $\{\mathbb{P}_\theta, \theta \in \Omega\}$, where θ represents some type of "parameter" to be estimated. As we discuss in the sequel, the space Ω of possible parameters can take various forms, including subsets of vectors, matrices, or—in the nonparametric setting to be discussed in Chapters 13 and 14—subsets of regression or density functions. Suppose that we observe a collection of n samples $Z_1^n = (Z_1, \ldots, Z_n)$, where each sample Z_i takes values in some space $\mathcal{Z}$, and is drawn independently according to some distribution $\mathbb{P}$. In the simplest setting, known as the well-specified case, the distribution $\mathbb{P}$ is a member of our parameterized family—say $\mathbb{P} = \mathbb{P}_{\theta^*}$—and our goal is to estimate the unknown parameter θ^*. However, our set-up will also allow for mis-specified models, in which case the target parameter θ^* is defined as the minimizer of the population cost function—in particular, see equation (9.2) below.

The first ingredient of a general M-estimator is a cost function $\mathcal{L}_n \colon \Omega \times \mathcal{Z}^n \to \mathbb{R}$, where the value $\mathcal{L}_n(\theta; Z_1^n)$ provides a measure of the fit of parameter θ to the data Z_1^n. Its expectation defines the *population cost function*—namely the quantity

$$\bar{\mathcal{L}}(\theta) := \mathbb{E}[\mathcal{L}_n(\theta; Z_1^n)]. \tag{9.1}$$

Implicit in this definition is that the expectation does not depend on the sample size n, a condition which holds in many settings (with appropriate scalings). For instance, it is often the case that the cost function has an additive decomposition of the form $\mathcal{L}_n(\theta; Z_1^n) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(\theta; Z_i)$, where $\mathcal{L} \colon \Omega \times \mathcal{Z} \to \mathbb{R}$ is the cost defined for a single sample. Of course, any likelihood-based cost function decomposes in this way when the samples are drawn in an independent and identically distributed manner, but such cost functions can also be useful for dependent data.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

Next we define the *target parameter* as the minimum of the population cost function

$$\theta^* = \arg\min_{\theta\in\Omega} \bar{\mathcal{L}}(\theta). \tag{9.2}$$

In many settings—in particular, when $\mathcal{L}_n$ is the negative log-likelihood of the data—this minimum is achieved at an interior point of Ω, in which case θ^* must satisfy the zero-gradient equation $\nabla\bar{\mathcal{L}}(\theta^*) = 0$. However, we do not assume this condition in our general analysis.

With this set-up, our goal is to estimate θ^* on the basis of the observed samples $Z_1^n = \{Z_1, \ldots, Z_n\}$. In order to do so, we combine the empirical cost function with a regularizer or penalty function $\Phi\colon \Omega \to \mathbb{R}$. As will be clarified momentarily, the purpose of this regularizer is to enforce a certain type of structure expected in θ^*. Our overall estimator is based on solving the optimization problem

$$\widehat{\theta} \in \arg\min_{\theta\in\Omega} \{\mathcal{L}_n(\theta; Z_1^n) + \lambda_n\Phi(\theta)\}, \tag{9.3}$$

where $\lambda_n > 0$ is a user-defined regularization weight. The estimator (9.3) is known as an *M-estimator*, where the “M” stands for minimization (or maximization).

Remark: An important remark on notation is needed before proceeding. From here onwards, we will frequently adopt $\mathcal{L}_n(\theta)$ as a shorthand for $\mathcal{L}_n(\theta; Z_1^n)$, remembering that the subscript n reflects implicitly the dependence on the underlying samples. We also adopt the same notation for the derivatives of the empirical cost function.

Let us illustrate this set-up with some examples.

Example 9.1 (Linear regression and Lasso) We begin with the problem of linear regression previously studied in Chapter 7. In this case, each sample takes the form $Z_i = (x_i, y_i)$, where $x_i \in \mathbb{R}^d$ is a covariate vector, and $y_i \in \mathbb{R}$ is a response variable. In the simplest case, we assume that the data are generated exactly from a linear model, so that $y_i = \langle x_i, \theta^*\rangle + w_i$, where w_i is some type of stochastic noise variable, assumed to be independent of x_i. The least-squares estimator is based on the quadratic cost function

$$\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(y_i - \langle x_i, \theta\rangle)^2 = \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2,$$

where we recall from Chapter 7 our usual notation for the vector $y \in \mathbb{R}^n$ of response variables and design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$. When the response–covariate pairs (y_i, x_i) are drawn from a linear model with regression vector θ^*, then the population cost function takes the form

$$\mathbb{E}_{x,y}[\frac{1}{2}(y - \langle x, \theta\rangle)^2] = \frac{1}{2}(\theta - \theta^*)^{\mathrm{T}}\mathbf{\Sigma}(\theta - \theta^*) + \frac{1}{2}\sigma^2 = \frac{1}{2}\|\sqrt{\mathbf{\Sigma}}\,(\theta - \theta^*)\|_2^2 + \frac{1}{2}\sigma^2,$$

where $\mathbf{\Sigma} := \mathrm{cov}(x_1)$ and $\sigma^2 := \mathrm{var}(w_1)$. Even when the samples are not drawn from a linear model, we can still define θ^* as a minimizer of the population cost function $\theta \mapsto \mathbb{E}_{x,y}[(y - \langle x, \theta\rangle)^2]$. In this case, the linear function $x \mapsto \langle x, \theta^*\rangle$ provides the best linear approximation of the regression function $x \mapsto \mathbb{E}[y \mid x]$.

As discussed in Chapter 7, there are many cases in which the target regression vector θ^*

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

is expected to be sparse, and in such settings, a good choice of regularizer Φ is the ℓ_1-norm $\Phi(\theta) = \sum_{j=1}^{d} |\theta_j|$. In conjunction with the least-squares loss, we obtain the Lasso estimator

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle x_i,\, \theta\rangle)^2 + \lambda_n \sum_{j=1}^{d} |\theta_j| \right\} \tag{9.4}$$

as a special case of the general estimator (9.3). See Chapter 7 for an in-depth analysis of this particular M-estimator. ♣

As our first extension of the basic Lasso (9.4), we now consider a more general family of regression problems.

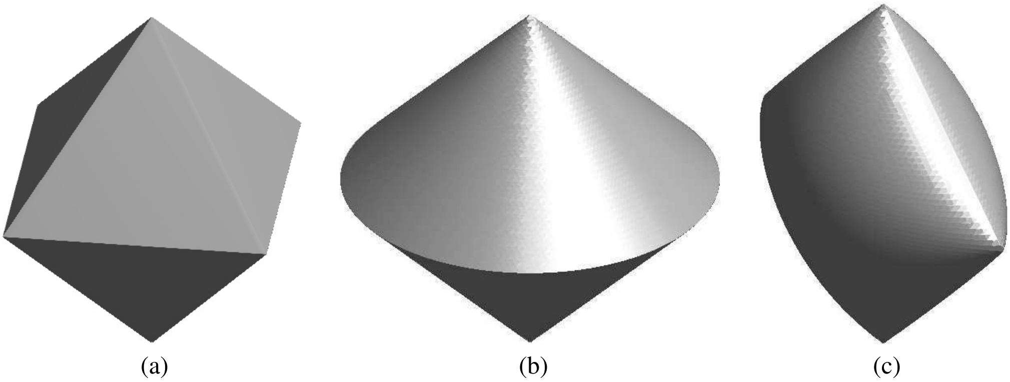


Figure 9.1 Illustration of unit balls of different norms in $\mathbb{R}^3$. (a) The ℓ_1-ball generated by $\Phi(\theta) = \sum_{j=1}^{3} |\theta_j|$. (b) The group Lasso ball generated by $\Phi(\theta) = \sqrt{\theta_1^2 + \theta_2^2} + |\theta_3|$. (c) A group Lasso ball with overlapping groups, generated by $\Phi(\theta) = \sqrt{\theta_1^2 + \theta_2^2} + \sqrt{\theta_1^2 + \theta_3^2}$.

Example 9.2 (Generalized linear models and ℓ_1-regularization) We again consider samples of the form $Z_i = (x_i, y_i)$ where $x_i \in \mathbb{R}^d$ is a vector of covariates, but now the response variable y_i is allowed to take values in an arbitrary space $\mathcal{Y}$. The previous example of linear regression corresponds to the case $\mathcal{Y} = \mathbb{R}$. A different example is the problem of binary classification, in which the response y_i represents a class label belonging to $\mathcal{Y} = \{0, 1\}$. For applications that involve responses that take on non-negative integer values—for instance, photon counts in imaging applications—the choice $\mathcal{Y} = \{0, 1, 2, \ldots\}$ is appropriate.

The family of *generalized linear models*, or GLMs for short, provides a unified approach to these different types of regression problems. Any GLM is based on modeling the conditional distribution of the response $y \in \mathcal{Y}$ given the covariate $x \in \mathbb{R}^d$ in an exponential family form, namely as

$$\mathbb{P}_{\theta^*}(y \mid x) = h_\sigma(y)\, \exp\left\{ \frac{y\,\langle x,\, \theta^*\rangle - \psi(\langle x,\, \theta^*\rangle)}{c(\sigma)} \right\}, \tag{9.5}$$

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

where $c(\sigma)$ is a scale parameter, and the function $\psi\colon \mathbb{R} \to \mathbb{R}$ is the partition function of the underlying exponential family.

Many standard models are special cases of the generalized linear family (9.5). First, consider the standard linear model $y = \langle x,\, \theta^* \rangle + w$, where $w \sim \mathcal{N}(0, \sigma^2)$. Setting $c(\sigma) = \sigma^2$ and $\psi(t) = t^2/2$, the conditional distribution (9.5) corresponds to that of a $\mathcal{N}(\langle x,\, \theta^* \rangle, \sigma^2)$ variate, as required. Similarly, in the logistic model for binary classification, we assume that the log-odds ratio is given by $\langle x,\, \theta^* \rangle$—that is,

$$\log \frac{\mathbb{P}_{\theta^*}(y = 1 \mid x)}{\mathbb{P}_{\theta^*}(y = 0 \mid x)} = \langle x,\, \theta^* \rangle. \tag{9.6}$$

This assumption again leads to a special case of the generalized linear model (9.5), this time with $c(\sigma) \equiv 1$ and $\psi(t) = \log(1 + \exp(t))$. As a final example, when the response $y \in \{0, 1, 2, \ldots\}$ represents some type of count, it can be appropriate to model y as conditionally Poisson with mean $\mu = e^{\langle x, \theta^* \rangle}$. This assumption leads to a generalized linear model (9.5) with $\psi(t) = \exp(t)$ and $c(\sigma) \equiv 1$. See Exercise 9.3 for verification of these properties.

Given n samples from the model (9.5), the negative log-likelihood takes the form

$$\mathcal{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \psi(\langle x_i,\, \theta \rangle) - \left\langle \frac{1}{n} \sum_{i=1}^{n} y_i x_i,\, \theta \right\rangle. \tag{9.7}$$

Here we have rescaled the log-likelihood by $1/n$ for later convenience, and also dropped the scale factor $c(\sigma)$, since it is independent of θ. When the true regression vector θ^* is expected to be sparse, then it is again reasonable to use the ℓ_1-norm as a regularizer, and combining with the cost function (9.7) leads to the *generalized linear Lasso*

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \psi(\langle x_i,\, \theta \rangle) - \left\langle \frac{1}{n} \sum_{i=1}^{n} y_i x_i,\, \theta \right\rangle + \lambda_n \|\theta\|_1 \right\}. \tag{9.8}$$

When $\psi(t) = t^2/2$, this objective function is equivalent to the standard Lasso, apart from the constant term $\frac{1}{2n} \sum_{i=1}^{n} y_i^2$ that has no effect on $\widehat{\theta}$. ♣

Thus far, we have discussed only the ℓ_1-norm. There are various extensions of the ℓ_1-norm that are based on some type of grouping of the coefficients.

Example 9.3 (Group Lasso) Let $\mathcal{G} = \{g_1, \ldots, g_T\}$ be a disjoint partition of the index set $\{1, \ldots, d\}$—that is, each group g_j is a subset of the index set, disjoint from every other group, and the union of all T groups covers the full index set. See panel (a) in Figure 9.3 for an example of a collection of overlapping groups.

For a given vector $\theta \in \mathbb{R}^d$, we let θ_g denote the d-dimensional vector with components equal to θ on indices within g, and zero in all other positions. For a given base norm $\|\cdot\|$, we then define the *group Lasso norm*

$$\Phi(\theta) := \sum_{g \in \mathcal{G}} \|\theta_g\|. \tag{9.9}$$

The standard form of the group Lasso uses the ℓ_2-norm as the base norm, so that we obtain a block ℓ_1/ℓ_2-norm—namely, the ℓ_1-norm of the ℓ_2-norms within each group. See Figure 9.1(b) for an illustration of the norm (9.9) with the blocks $g_1 = \{1, 2\}$ and $g_2 = \{3\}$. The

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

block ℓ_1/ℓ_∞-version of the group Lasso has also been studied extensively. Apart from the basic group Lasso (9.9), another variant involves associating a positive weight ω_g with each group. ♣

In the preceding example, the groups were non-overlapping. The same regularizer (9.9) can also be used in the case of overlapping groups; it remains a norm as long as the groups cover the space. For instance, Figure 9.1(c) shows the unit ball generated by the overlapping groups $g_1 = \{1,2\}$ and $g_2 = \{1,3\}$ in $\mathbb{R}^3$. However, the standard group Lasso (9.9) with overlapping groups has a property that can be undesirable. Recall that the motivation for group-structured penalties is to estimate parameter vectors whose support lies within a union of a (relatively small) subset of groups. However, when used as a regularizer in an M-estimator, the standard group Lasso (9.9) with overlapping groups typically leads to solutions with support contained in the *complement* of a union of groups. For instance, in the example shown in Figure 9.1(c) with groups $g_1 = \{1,2\}$ and $g_2 = \{1,3\}$, apart from the all-zero solution that has empty support set, or a solution with the complete support $\{1,2,3\}$, the penalty encourages solutions with supports equal to either $g_1^c = \{3\}$ or $g_2^c = \{2\}$.

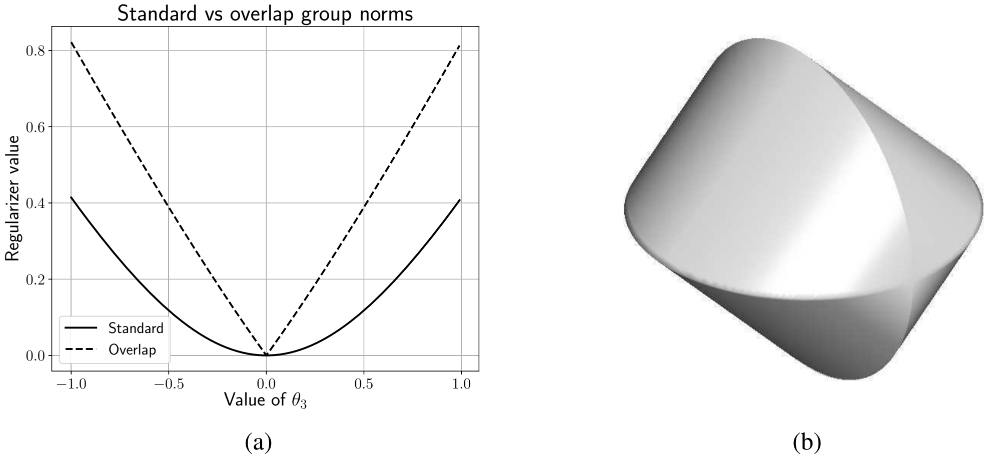


Figure 9.2 (a) Plots of the residual penalty $f(\theta_3) = \Phi(1,1,\theta_3) - \Phi(1,1,0)$ for the standard group Lasso (9.9) with a solid line and overlap group Lasso (9.10) with a dashed line, in the case of the groups $g_1 = \{1,2\}$ and $g_2 = \{1,3\}$. (b) Plot of the unit ball of the overlapping group Lasso norm (9.10) for the same groups as in panel (a).

Why is this the case? In the example given above, consider a vector $\theta \in \mathbb{R}^3$ such that θ_1, a variable shared by both groups, is active. For concreteness, say that $\theta_1 = \theta_2 = 1$, and consider the residual penalty $f(\theta_3) := \Phi(1,1,\theta_3) - \Phi(1,1,0)$ on the third coefficient. It takes the form

$$f(\theta_3) = \|(1,1)\|_2 + \|(1,\theta_3)\|_2 - \|(1,1)\|_2 - \|(1,0)\|_2 = \sqrt{1+\theta_3^2} - 1.$$

As shown by the solid curve in Figure 9.2(a), the function f is differentiable at $\theta_3 = 0$. Indeed, since $f'(\theta_3)\big|_{\theta_3=0} = 0$, this penalty *does not* encourage sparsity of the third coefficient.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

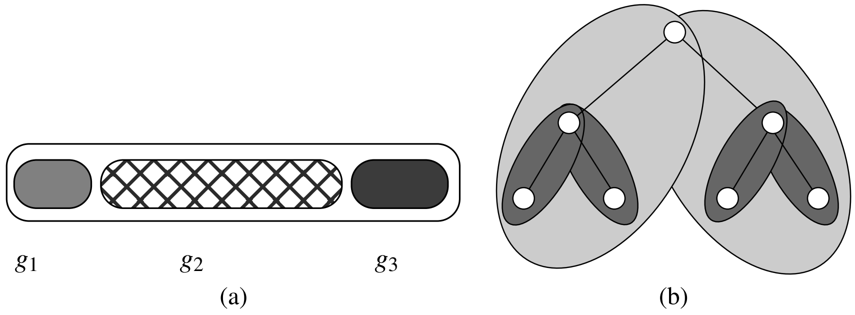


Figure 9.3 (a) Group Lasso penalty with non-overlapping groups. The groups $\{g_1, g_2, g_3\}$ form a disjoint partition of the index set $\{1, 2, \ldots, d\}$. (b) A total of $d = 7$ variables are associated with the vertices of a binary tree, and sub-trees are used to define a set of overlapping groups. Such overlapping group structures arise naturally in multiscale signal analysis.

A similar argument applies with the roles of θ_2 and θ_3 reversed. Consequently, if the shared first variable is active in an optimal solution, it is usually the case that the second and third variables will also be active, leading to a fully dense solution. See the bibliographic discussion for references that discuss this phenomenon in greater detail.

The overlapping group Lasso is a closely related but different penalty that is designed to overcome this potentially troublesome issue.

Example 9.4 (Overlapping group Lasso) As in Example 9.3, consider a collection of groups $\mathcal{G} = \{g_1, \ldots, g_T\}$, where each group is a subset of the index set $\{1, \ldots, d\}$. We require that the union over all groups covers the full index set, but we allow for overlaps among the groups. See panel (b) in Figure 9.3 for an example of a collection of overlapping groups.

When there actually is overlap, any vector θ has many possible group representations, meaning collections $\{w_g,\ g \in \mathcal{G}\}$ such that $\sum_{g \in \mathcal{G}} w_g = \theta$. The *overlap group norm* is based on minimizing over all such representations, as follows:

$$\Phi_{\text{over}}(\theta) := \inf_{\substack{\theta = \sum_{g \in \mathcal{G}} w_g \\ w_g,\ g \in \mathcal{G}}} \left\{ \sum_{g \in \mathcal{G}} \|w_g\| \right\}. \tag{9.10}$$

As we verify in Exercise 9.1, the variational representation (9.10) defines a valid norm on $\mathbb{R}^d$. Of course, when the groups are non-overlapping, this definition reduces to the previous one (9.9). Figure 9.2(b) shows the overlapping group norm (9.10) in the special case of the groups $g_1 = \{1, 2\}$ and $g_2 = \{1, 3\}$. Notice how it differs from the standard group Lasso (9.9) with the same choice of groups, as shown in Figure 9.1(c). ♣

When used as a regularizer in the general M-estimator (9.3), the overlapping group Lasso (9.10) tends to induce solution vectors with their support contained within a union of the groups. To understand this issue, let us return to the group set $g_1 = \{1, 2\}$ and $g_2 = \{1, 3\}$, and suppose once again that the first two variables are active, say $\theta_1 = \theta_2 = 1$. The residual penalty on θ_3 then takes the form

$$f_{\text{over}}(\theta_3) := \Phi_{\text{over}}(1,1,\theta_3) - \Phi_{\text{over}}(1,1,0) = \inf_{\alpha\in\mathbb{R}} \{\|(\alpha,1)\|_2 + \|(1-\alpha,\theta_3)\|_2\} - \sqrt{2}.$$

It can be shown that this function behaves like the ℓ_1-norm around the origin, so that it tends to encourage sparsity in θ_3. See Figure 9.2(b) for an illustration.

Up to this point, we have considered vector estimation problems, in which the parameter space Ω is some subspace of $\mathbb{R}^d$. We now turn to various types of matrix estimation problems, in which the parameter space is some subset of $\mathbb{R}^{d_1\times d_2}$, the space of all $(d_1 \times d_2)$-dimensional matrices. Of course, any such problem can be viewed as a vector estimation problem, simply by transforming the matrix to a $D = d_1 d_2$ vector. However, it is often more natural to retain the matrix structure of the problem. Let us consider some examples.

Example 9.5 (Estimation of Gaussian graphical models) Any zero-mean Gaussian random vector with a strictly positive definite covariance matrix $\mathbf{\Sigma} \succ 0$ has a density of the form

$$\mathbb{P}(x_1,\ldots,x_d;\mathbf{\Theta}^*) \propto \sqrt{\det(\mathbf{\Theta}^*)}\, e^{-\frac{1}{2}x^{\mathrm{T}}\mathbf{\Theta}^* x}, \tag{9.11}$$

where $\mathbf{\Theta}^* = (\mathbf{\Sigma})^{-1}$ is the inverse covariance matrix, also known as the precision matrix. In many cases, the components of the random vector $X = (X_1,\ldots,X_d)$ satisfy various types of conditional independence relationships: for instance, it might be the case that X_j is conditionally independent of X_k given the other variables $X_{\setminus\{j,k\}}$. In the Gaussian case, it is a consequence of the Hammersley–Clifford theorem that this conditional independence statement holds if and only if the precision matrix $\mathbf{\Theta}^*$ has a zero in position (j,k). Thus, conditional independence is directly captured by the sparsity of the precision matrix. See Chapter 11 for further details on this relationship between conditional independence, and the structure of $\mathbf{\Theta}^*$.

Given a Gaussian model that satisfies many conditional independence relationships, the precision matrix will be sparse, in which case it is natural to use the elementwise ℓ_1-norm $\Phi(\mathbf{\Theta}) = \sum_{j\neq k} |\Theta_{jk}|$ as a regularizer. Here we have chosen not to regularize the diagonal entries, since they all must be non-zero so as to ensure strict positive definiteness. Combining this form of ℓ_1-regularization with the Gaussian log-likelihood leads to the estimator

$$\widehat{\mathbf{\Theta}} \in \arg\min_{\mathbf{\Theta}\in\mathcal{S}^{d\times d}} \left\{ \langle\!\langle \mathbf{\Theta},\, \widehat{\mathbf{\Sigma}} \rangle\!\rangle - \log\det\mathbf{\Theta} + \lambda_n \sum_{j\neq k} |\Theta_{jk}| \right\}, \tag{9.12}$$

where $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n x_i x_i^{\mathrm{T}}$ is the sample covariance matrix. This combination corresponds to another special case of the general estimator (9.3), known as the *graphical Lasso*, which we analyze in Chapter 11. ♣

The problem of multivariate regression is a natural extension of a standard regression problem, which involves scalar response variables, to the vector-valued setting.

Example 9.6 (Multivariate regression) In a multivariate regression problem, we observe samples of the form $(z_i, y_i) \in \mathbb{R}^p \times \mathbb{R}^T$, and our goal is to use the vector of features z_i to

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

predict the vector of responses $y_i \in \mathbb{R}^T$. Let $\mathbf{Y} \in \mathbb{R}^{n \times T}$ and $\mathbf{Z} \in \mathbb{R}^{n \times p}$ be matrices with y_i and z_i, respectively, as their ith row. In the simplest case, we assume that the response matrix $\mathbf{Y}$ and covariate matrix $\mathbf{Z}$ are linked via the linear model

$$\mathbf{Y} = \mathbf{Z}\mathbf{\Theta}^* + \mathbf{W}, \tag{9.13}$$

where $\mathbf{\Theta}^* \in \mathbb{R}^{p \times T}$ is a matrix of regression coefficients, and $\mathbf{W} \in \mathbb{R}^{n \times T}$ is a stochastic noise matrix. See Figure 9.4 for an illustration.

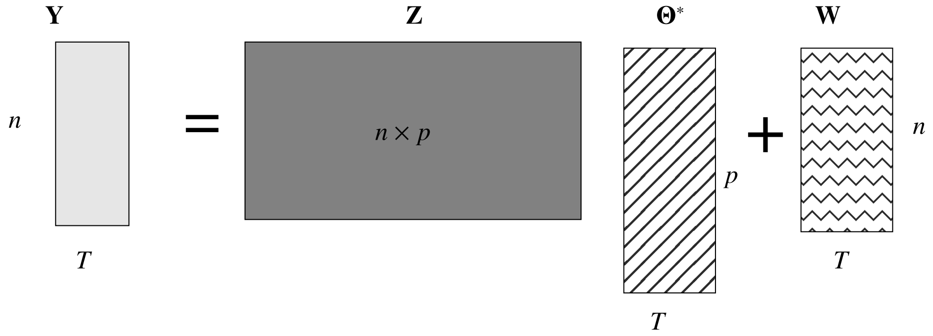


Figure 9.4 Illustration of the multivariate linear regression model: a data set of n observations consists of a matrix $\mathbf{Y} \in \mathbb{R}^{n \times T}$ of multivariate responses, and a matrix $\mathbf{Z} \in \mathbb{R}^{n \times p}$ of covariates, in this case shared across the tasks. Our goal is to estimate the matrix $\mathbf{\Theta}^* \in \mathbb{R}^{p \times T}$ of regression coefficients.

One way in which to view the model (9.13) is as a collection of T different p-dimensional regression problems of the form

$$Y_{\cdot,t} = \mathbf{Z}\Theta^*_{\cdot,t} + W_{\cdot,t}, \qquad \text{for } t = 1, \ldots, T,$$

where $Y_{\cdot,t} \in \mathbb{R}^n$, $\Theta^*_{\cdot,t} \in \mathbb{R}^p$ and $W_{\cdot,t} \in \mathbb{R}^n$ are the tth columns of the matrices $\mathbf{Y}$, $\mathbf{\Theta}^*$ and $\mathbf{W}$, respectively. One could then estimate each column $\Theta^*_{\cdot,t}$ separately by solving a standard univariate regression problem.

However, many applications lead to interactions between the different columns of $\mathbf{\Theta}^*$, which motivates solving the univariate regression problems in a joint manner. For instance, it is often the case that there is a subset of features—that is, a subset of the rows of $\mathbf{\Theta}^*$—that are relevant for prediction in all T regression problems. For estimating such a row-sparse matrix, a natural regularizer is the row-wise $(2,1)$-norm $\Phi(\mathbf{\Theta}) := \sum_{j=1}^p \|\Theta_{j,\cdot}\|_2$, where $\Theta_{j,\cdot} \in \mathbb{R}^T$ denotes the jth row of the matrix $\mathbf{\Theta} \in \mathbb{R}^{p \times T}$. Note that this regularizer is a special case of the general group penalty (9.9). Combining this regularizer with the least-squares cost, we obtain

$$\widehat{\mathbf{\Theta}} \in \arg \min_{\mathbf{\Theta} \in \mathbb{R}^{p \times T}} \left\{ \frac{1}{2n} |||\mathbf{Y} - \mathbf{Z}\mathbf{\Theta}|||_{\mathrm{F}}^2 + \lambda_n \sum_{j=1}^p \|\Theta_{j,\cdot}\|_2 \right\}. \tag{9.14}$$

This estimator is often referred to as the *multivariate group Lasso*, for obvious reasons. The underlying optimization problem is an instance of a second-order cone problem (SOCP), and can be solved efficiently by a variety of algorithms; see the bibliography section for further discussion. ♣

Other types of structure are also possible in multivariate regression problems, and lead to different types of regularization.

Example 9.7 (Overlapping group Lasso and multivariate regression) There is an interesting extension of the row-sparse model from Example 9.6, one which leads to an instance of the overlapping group Lasso (9.10). The row-sparse model assumes that there is a relatively small subset of predictors, each of which is active in *all* of the T tasks. A more flexible model allows for the possibility of a subset of predictors that are shared among all tasks, coupled with a subset of predictors that appear in only one (or relatively few) tasks. This type of structure can be modeled by decomposing the regression matrix $\mathbf{\Theta}^*$ as the sum of a row-sparse matrix $\mathbf{\Omega}^*$ along with an elementwise-sparse matrix $\mathbf{\Gamma}^*$. If we impose a group $\ell_{1,2}$-norm on the row-sparse component and an ordinary ℓ_1-norm on the element-sparse component, then we are led to the estimator

$$(\widehat{\mathbf{\Omega}}, \widehat{\mathbf{\Gamma}}) \in \arg \min_{\mathbf{\Omega},\mathbf{\Gamma} \in \mathbb{R}^{d \times T}} \left\{ \frac{1}{2n} |||\mathbf{Y} - \mathbf{Z}(\mathbf{\Omega} + \mathbf{\Gamma})|||_{\mathrm{F}}^2 + \lambda_n \sum_{j=1}^{d} \|\Omega_{j,\cdot}\|_2 + \mu_n \|\mathbf{\Gamma}\|_1 \right\}, \tag{9.15}$$

where $\lambda_n, \mu_n > 0$ are regularization parameters to be chosen. Any solution to this optimization problem defines an estimate of the full regression matrix via $\widehat{\mathbf{\Theta}} = \widehat{\mathbf{\Omega}} + \widehat{\mathbf{\Gamma}}$.

We have defined the estimator (9.15) as an optimization problem over the matrix pair $(\mathbf{\Omega}, \mathbf{\Gamma})$, using a separate regularizer for each matrix component. Alternatively, we can formulate it as a direct estimator for $\widehat{\mathbf{\Theta}}$. In particular, by making the substitution $\mathbf{\Theta} = \mathbf{\Omega} + \mathbf{\Gamma}$, and minimizing over both $\mathbf{\Theta}$ and the pair $(\mathbf{\Omega}, \mathbf{\Gamma})$ subject to this linear constraint, we obtain the equivalent formulation

$$\widehat{\mathbf{\Theta}} \in \arg \min_{\mathbf{\Theta} \in \mathbb{R}^{d \times T}} \left\{ \frac{1}{2n} |||\mathbf{Y} - \mathbf{Z}\mathbf{\Theta}|||_{\mathrm{F}}^2 + \lambda_n \underbrace{\left\{ \inf_{\mathbf{\Omega}+\mathbf{\Gamma}=\mathbf{\Theta}} \|\mathbf{\Omega}\|_{1,2} + \omega_n \|\mathbf{\Gamma}\|_1 \right\}}_{\Phi_{\mathrm{over}}(\mathbf{\Theta})} \right\}, \tag{9.16}$$

where $\omega_n = \frac{\mu_n}{\lambda_n}$. In this direct formulation, we see that the assumed decomposition leads to an interesting form of the overlapping group norm. We return to study the estimator (9.16) in Section 9.7. ♣

In other applications of multivariate regression, one might imagine that the individual regression vectors—that is, the columns $\Theta^*_{\cdot,t} \in \mathbb{R}^p$—all lie within some low-dimensional subspace, corresponding to some hidden meta-features, so that it has relatively low rank. Many other problems, to be discussed in more detail in Chapter 10, also lead to estimation problems that involve rank constraints. In such settings, the ideal approach would be to impose an explicit rank constraint within our estimation procedure. Unfortunately, when viewed as function on the space of $d_1 \times d_2$ matrices, the rank function is non-convex, so that this approach is not computationally feasible. Accordingly, we are motivated to study convex relaxations of rank constraints.

Example 9.8 (Nuclear norm as a relaxation of rank) The *nuclear norm* provides a natural relaxation of the rank of a matrix, one which is analogous to the ℓ_1-norm as a relaxation of

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

the cardinality of a vector. In order to define the nuclear norm, we first recall the *singular value decomposition*, or SVD for short, of a matrix $\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}$. Letting $d' = \min\{d_1, d_2\}$, the SVD takes the form

$$\mathbf{\Theta} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}, \tag{9.17}$$

where $\mathbf{U} \in \mathbb{R}^{d_1 \times d'}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times d'}$ are orthonormal matrices (meaning that $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_{d'}$). The matrix $\mathbf{D} \in \mathbb{R}^{d' \times d'}$ is diagonal with its entries corresponding to the singular values of $\mathbf{\Theta}$, denoted by

$$\sigma_1(\mathbf{\Theta}) \geq \sigma_2(\mathbf{\Theta}) \geq \sigma_3(\mathbf{\Theta}) \geq \cdots \geq \sigma_{d'}(\mathbf{\Theta}) \geq 0. \tag{9.18}$$

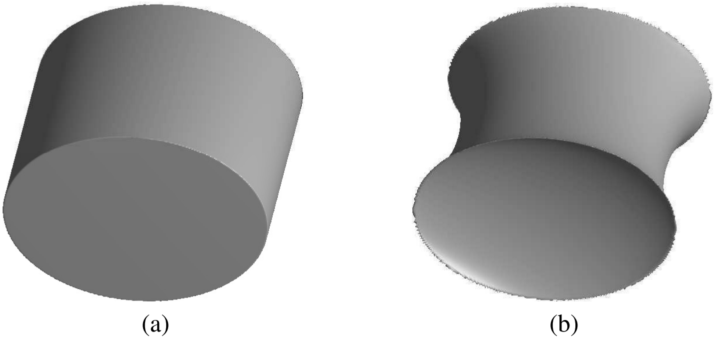


Figure 9.5 Illustration of the nuclear norm ball as a relaxation of a rank constraint. (a) Set of all matrices of the form $\mathbf{\Theta} = \begin{bmatrix} \theta_1 & \theta_2 \\ \theta_2 & \theta_3 \end{bmatrix}$ such that $|\!|\!|\mathbf{\Theta}|\!|\!|_{\mathrm{nuc}} \leq 1$. This is a projection of the unit ball of the nuclear norm ball onto the space of symmetric matrices. (b) For a parameter $q > 0$, the ℓ_q-"ball" of matrices is defined by $\mathbb{B}_q(1) = \{\mathbf{\Theta} \in \mathbb{R}^{2\times 2} \mid \sum_{j=1}^{2} \sigma_j(\mathbf{\Theta})^q \leq 1\}$. For all $q \in [0, 1)$, this is a non-convex set, and it is equivalent to the set of all rank-one matrices for $q = 0$.

Observe that the number of strictly positive singular values specifies the rank—that is, we have $\mathrm{rank}(\mathbf{\Theta}) = \sum_{j=1}^{d'} \mathbb{I}[\sigma_j(\mathbf{\Theta}) > 0]$. This observation, though not practically useful on its own, suggests a natural convex relaxation of a rank constraint, namely the *nuclear norm*

$$|\!|\!|\mathbf{\Theta}|\!|\!|_{\mathrm{nuc}} = \sum_{j=1}^{d'} \sigma_j(\mathbf{\Theta}), \tag{9.19}$$

corresponding to the ℓ_1-norm of the singular values.[1] As shown in Figure 9.5(a), the nuclear norm provides a convex relaxation of the set of low-rank matrices. ♣

There are a variety of other statistical models—in addition to multivariate regression—in which rank constraints play a role, and the nuclear norm relaxation is useful for many of them. These problems are discussed in detail in Chapter 10 to follow.

[1] No absolute value is necessary, since singular values are non-negative by definition.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

9.2 Decomposable regularizers and their utility

Having considered a general family of M-estimators (9.3) and illustrated it with various examples, we now turn to the development of techniques for bounding the estimation error $\widehat{\theta} - \theta^*$. The first ingredient in our analysis is a property of the regularizer known as decomposability. It is a geometric property, based on how the regularizer behaves over certain pairs of subspaces. The ℓ_1-norm is the canonical example of a decomposable norm, but various other norms also share this property. Decomposability implies that any optimum $\widehat{\theta}$ to the M-estimator (9.3) belongs to a very special set, as shown in Proposition 9.13.

From here onwards, we assume that the set Ω is endowed with an inner product $\langle \cdot, \cdot \rangle$, and we use $\|\cdot\|$ to denote the norm induced by this inner product. The standard examples to keep in mind are

- the space $\mathbb{R}^d$ with the usual Euclidean inner product, or more generally with a weighted Euclidean inner product, and
- the space $\mathbb{R}^{d_1 \times d_2}$ equipped with the trace inner product (10.1).

Given a vector $\theta \in \Omega$ and a subspace $\mathbb{S}$ of Ω, we use $\theta_{\mathbb{S}}$ to denote the projection of θ onto $\mathbb{S}$. More precisely, we have

$$\theta_{\mathbb{S}} := \arg\min_{\widetilde{\theta} \in \mathbb{S}} \|\widetilde{\theta} - \theta\|^2. \tag{9.20}$$

These projections play an important role in the sequel; see Exercise 9.2 for some examples.

9.2.1 Definition and some examples

The notion of a decomposable regularizer is defined in terms of a pair of subspaces $\mathbb{M} \subseteq \overline{\mathbb{M}}$ of $\mathbb{R}^d$. The role of the *model subspace* $\mathbb{M}$ is to capture the constraints specified by the model; for instance, as illustrated in the examples to follow, it might be the subspace of vectors with a particular support or a subspace of low-rank matrices. The orthogonal complement of the space $\overline{\mathbb{M}}$, namely the set

$$\overline{\mathbb{M}}^{\perp} := \left\{ v \in \mathbb{R}^d \mid \langle u, v \rangle = 0 \quad \text{for all } u \in \overline{\mathbb{M}} \right\}, \tag{9.21}$$

is referred to as the *perturbation subspace*, representing deviations away from the model subspace $\mathbb{M}$. In the ideal case, we have $\overline{\mathbb{M}}^{\perp} = \mathbb{M}^{\perp}$, but the definition allows for the possibility that $\overline{\mathbb{M}}$ is strictly larger than $\mathbb{M}$, so that $\overline{\mathbb{M}}^{\perp}$ is strictly smaller than $\mathbb{M}^{\perp}$. This generality is needed for treating the case of low-rank matrices and nuclear norm, as discussed in Chapter 10.

Definition 9.9 Given a pair of subspaces $\mathbb{M} \subseteq \overline{\mathbb{M}}$, a norm-based regularizer Φ is *decomposable* with respect to $(\mathbb{M}, \overline{\mathbb{M}}^{\perp})$ if

$$\Phi(\alpha + \beta) = \Phi(\alpha) + \Phi(\beta) \qquad \text{for all } \alpha \in \mathbb{M} \text{ and } \beta \in \overline{\mathbb{M}}^{\perp}. \tag{9.22}$$

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

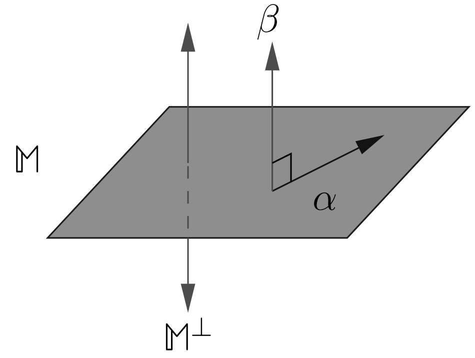


Figure 9.6 In the ideal case, decomposability is defined in terms of a subspace pair $(\mathbb{M}, \mathbb{M}^\perp)$. For any $\alpha \in \mathbb{M}$ and $\beta \in \mathbb{M}^\perp$, the regularizer should decompose as $\Phi(\alpha + \beta) = \Phi(\alpha) + \Phi(\beta)$.

See Figure 9.6 for the geometry of this definition. In order to build some intuition, let us consider the ideal case $\mathbb{M} = \overline{\mathbb{M}}$, so that the decomposition (9.22) holds for all pairs $(\alpha, \beta) \in \mathbb{M} \times \mathbb{M}^\perp$. For any given pair (α, β) of this form, the vector $\alpha + \beta$ can be interpreted as perturbation of the model vector α away from the subspace $\mathbb{M}$, and it is desirable that the regularizer penalize such deviations as much as possible. By the triangle inequality for a norm, we always have $\Phi(\alpha + \beta) \leq \Phi(\alpha) + \Phi(\beta)$, so that the decomposability condition (9.22) holds if and only if the triangle inequality is tight for all pairs $(\alpha, \beta) \in (\mathbb{M}, \mathbb{M}^\perp)$. It is exactly in this setting that the regularizer penalizes deviations away from the model subspace $\mathbb{M}$ as much as possible.

Let us consider some illustrative examples:

Example 9.10 (Decomposability and sparse vectors) We begin with the ℓ_1-norm, which is the canonical example of a decomposable regularizer. Let S be a given subset of the index set $\{1, \ldots, d\}$ and S^c be its complement. We then define the model subspace

$$\mathbb{M} \equiv \mathbb{M}(S) := \{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \in S^c\}, \tag{9.23}$$

corresponding to the set of all vectors that are supported on S. Observe that

$$\mathbb{M}^\perp(S) = \{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \in S\}.$$

With these definitions, it is then easily seen that for any pair of vectors $\alpha \in \mathbb{M}(S)$ and $\beta \in \mathbb{M}^\perp(S)$, we have

$$\|\alpha + \beta\|_1 = \|\alpha\|_1 + \|\beta\|_1,$$

showing that the ℓ_1-norm is decomposable with respect to the pair $(\mathbb{M}(S), \mathbb{M}^\perp(S))$. ♣

Example 9.11 (Decomposability and group sparse norms) We now turn to the notion of decomposability for the group Lasso norm (9.9). In this case, the subspaces are defined in terms of subsets of groups. More precisely, given any subset $S_{\mathcal{G}} \subset \mathcal{G}$ of the group index set, consider the set

$$\mathbb{M}(S_{\mathcal{G}}) := \{\theta \in \Omega \mid \theta_g = 0 \quad \text{for all } g \notin S_{\mathcal{G}}\}, \tag{9.24}$$

corresponding to the subspace of vectors supported only on groups indexed by $S_{\mathcal{G}}$. Note that the orthogonal subspace is given by $\mathbb{M}^{\perp}(S_{\mathcal{G}}) = \{\theta \in \Omega \mid \theta_g = 0 \text{ for all } g \in S_{\mathcal{G}}\}$. Letting $\alpha \in \mathbb{M}(S_{\mathcal{G}})$ and $\beta \in \mathbb{M}^{\perp}(S_{\mathcal{G}})$ be arbitrary, we have

$$\Phi(\alpha + \beta) = \sum_{g \in S_{\mathcal{G}}} \|\alpha_g\| + \sum_{g \in S_{\mathcal{G}}^c} \|\beta_g\| = \Phi(\alpha) + \Phi(\beta),$$

thus showing that the group norm is decomposable with respect to the pair $(\mathbb{M}(S_{\mathcal{G}}), \mathbb{M}^{\perp}(S_{\mathcal{G}}))$.
♣

In the preceding example, we considered the case of non-overlapping groups. It is natural to ask whether the same decomposability—that is, with respect to the pair $(\mathbb{M}(S_{\mathcal{G}}), \mathbb{M}^{\perp}(S_{\mathcal{G}}))$—continues to hold for the ordinary group Lasso $\|\theta\|_{\mathcal{G}} = \sum_{g \in \mathcal{G}} \|\theta_g\|$ when the groups are allowed to be overlapping. A little thought shows that this is not the case in general: for instance, in the case $\theta \in \mathbb{R}^4$, consider the overlapping groups $g_1 = \{1, 2\}$, $g_2 = \{2, 3\}$ and $g_3 = \{3, 4\}$. If we let $S_{\mathcal{G}} = \{g_1\}$, then

$$\mathbb{M}^{\perp}(S_{\mathcal{G}}) = \{\theta \in \mathbb{R}^4 \mid \theta_1 = \theta_2 = 0\}.$$

The vector $\alpha = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$ belongs to $\mathbb{M}(S_{\mathcal{G}})$, and the vector $\beta = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$ belongs to $\mathbb{M}^{\perp}(S_{\mathcal{G}})$. In the case of the group ℓ_1/ℓ_2-norm $\|\theta\|_{\mathcal{G},2} = \sum_{g \in \mathcal{G}} \|\theta_g\|_2$, we have $\|\alpha + \beta\|_{\mathcal{G},2} = 1 + \sqrt{2} + 1$, but

$$\|\alpha\|_{\mathcal{G},2} + \|\beta\|_{\mathcal{G},2} = 1 + 1 + 1 + 1 = 4 \ > \ 2 + \sqrt{2}, \tag{9.25}$$

showing that decomposability is violated. However, this issue can be addressed by a different choice of subspace pair, one that makes use of the additional freedom provided by allowing for $\overline{\mathbb{M}} \supsetneq \mathbb{M}$. We illustrate this procedure in the following:

Example 9.12 (Decomposability of ordinary group Lasso with overlapping groups) As before, let $S_{\mathcal{G}}$ be a subset of the group index set $\mathcal{G}$, and define the subspace $\mathbb{M}(S_{\mathcal{G}})$. We then define the augmented group set

$$\widetilde{S}_{\mathcal{G}} := \Big\{ g \in \mathcal{G} \mid g \cap \bigcup_{h \in S_{\mathcal{G}}} h \neq \emptyset \Big\}, \tag{9.26}$$

corresponding to the set of groups with non-empty intersection with some group in $S_{\mathcal{G}}$. Note that in the case of non-overlapping groups, we have $\widetilde{S}_{\mathcal{G}} = S_{\mathcal{G}}$, whereas $\widetilde{S}_{\mathcal{G}} \supseteq S_{\mathcal{G}}$ in the more general case of overlapping groups. This augmented set defines the subspace $\overline{\mathbb{M}} := \mathbb{M}(\widetilde{S}_{\mathcal{G}}) \supseteq \mathbb{M}(S_{\mathcal{G}})$, and we claim that the overlapping group norm is decomposable with respect to the pair $(\mathbb{M}(S_{\mathcal{G}}), \mathbb{M}^{\perp}(\widetilde{S}_{\mathcal{G}}))$.

Indeed, let α and β be arbitrary members of $\mathbb{M}(S_{\mathcal{G}})$ and $\mathbb{M}^{\perp}(\widetilde{S}_{\mathcal{G}})$, respectively. Note that any element of $\mathbb{M}^{\perp}(\widetilde{S}_{\mathcal{G}})$ can have support only on the subset $\bigcup_{h \notin \widetilde{S}_{\mathcal{G}}} h$; at the same time, this subset has no overlap with $\bigcup_{g \in S_{\mathcal{G}}} g$, and any element of $\mathbb{M}(S_{\mathcal{G}})$ is supported on this latter

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

subset. As a consequence of these properties, we have

$$\|\alpha + \beta\|_{\mathcal{G}} = \sum_{g \in \mathcal{G}} (\alpha + \beta)_g = \sum_{g \in \widetilde{S}_{\mathcal{G}}} \alpha_g + \sum_{g \notin \widetilde{S}_{\mathcal{G}}} \beta_g = \|\alpha\|_{\mathcal{G}} + \|\beta\|_{\mathcal{G}},$$

as claimed. ♣

It is worthwhile observing how our earlier counterexample (9.25) is excluded by the construction given in Example 9.12. With the groups $g_1 = \{1, 2\}$, $g_2 = \{2, 3\}$ and $g_3 = \{3, 4\}$, combined with the subset $S_{\mathcal{G}} = \{g_1\}$, we have $\widetilde{S}_{\mathcal{G}} = \{g_1, g_2\}$. The vector $\beta = \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$ belongs to the subspace

$$\mathbb{M}^{\perp}(S_{\mathcal{G}}) = \{\theta \in \mathbb{R}^d \mid \theta_1 = \theta_2 = 0\},$$

but it does *not* belong to the smaller subspace

$$\mathbb{M}^{\perp}(\widetilde{S}_{\mathcal{G}}) = \{\theta \in \mathbb{R}^4 \mid \theta_1 = \theta_2 = \theta_3 = 0\}.$$

Consequently, it does not violate the decomposability property. However, note that there is a statistical price to be paid by enlarging to the augmented set $\mathbb{M}(\widetilde{S}_{\mathcal{G}})$: as our later results demonstrate, the statistical estimation error scales as a function of the size of this set.

As discussed previously, many problems involve estimating low-rank matrices, in which context the nuclear norm (9.19) plays an important role. In Chapter 10, we show how the nuclear norm is decomposable with respect to appropriately chosen subspaces. Unlike our previous examples (in which $\mathbb{M} = \overline{\mathbb{M}}$), in this case we need to use the full flexibility of our definition, and choose $\overline{\mathbb{M}}$ to be a strict superset of $\mathbb{M}$.

Finally, it is worth noting that sums of decomposable regularizers over disjoint sets of parameters remain decomposable: that is, if Φ_1 and Φ_2 are decomposable with respect to subspaces over Ω_1 and Ω_2 respectively, then the sum $\Phi_1 + \Phi_2$ remains decomposable with respect to the same subspaces extended to the Cartesian product space $\Omega_1 \times \Omega_2$. For instance, this property is useful for the matrix decomposition problems discussed in Chapter 10, which involve a pair of matrices $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$, and the associated regularizers $\Phi_1(\mathbf{\Lambda}) = |||\mathbf{\Lambda}|||_{\text{nuc}}$ and $\Phi_2(\mathbf{\Gamma}) = \|\mathbf{\Gamma}\|_1$.

9.2.2 A key consequence of decomposability

Why is decomposability important in the context of M-estimation? Ultimately, our goal is to provide bounds on the error vector $\widehat{\Delta} := \widehat{\theta} - \theta^*$ between any global optimum of the optimization problem (9.3) and the unknown parameter θ^*. In this section, we show that decomposability—in conjunction with a suitable choice for the regularization weight λ_n—ensures that the error $\widehat{\Delta}$ must lie in a very restricted set.

In order to specify a "suitable" choice of regularization parameter λ_n, we need to define the notion of the dual norm associated with our regularizer. Given any norm $\Phi\colon \mathbb{R}^d \to \mathbb{R}$, its dual norm is defined in a variational manner as

$$\Phi^*(v) := \sup_{\Phi(u) \le 1} \langle u,\, v \rangle \,. \tag{9.27}$$

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

Regularizer Φ	Dual norm Φ^*
ℓ_1-norm $\quad \Phi(u) = \sum_{j=1}^d \lvert u_j \rvert$	ℓ_∞-norm $\quad \Phi^*(v) = \lVert v \rVert_\infty = \max\limits_{j=1,\ldots,d} \lvert v_j \rvert$
Group ℓ_1/ℓ_p-norm $\quad \Phi(u) = \sum_{g \in \mathcal{G}} \lVert u_g \rVert_p$ Non-overlapping groups	Group ℓ_∞/ℓ_q-norm $\quad \Phi^*(v) = \max\limits_{g \in \mathcal{G}} \lVert v_g \rVert_q$ $\frac{1}{p} + \frac{1}{q} = 1$
Nuclear norm $\quad \Phi(\mathbf{M}) = \sum\limits_{j=1}^{d} \sigma_j(\mathbf{M})$	ℓ_2-operator norm $\quad \Phi^*(\mathbf{N}) = \max\limits_{j=1,\ldots,d} \sigma_j(\mathbf{N})$ $d = \min\{d_1, d_2\}$
Overlap group norm $\Phi(u) = \inf\limits_{u = \sum_{g \in \mathcal{G}} w_g} \lVert w_g \rVert_p$	Overlap dual norm $\Phi^*(v) = \max_{g \in \mathcal{G}} \lVert v_g \rVert_q$
Sparse-low-rank decomposition norm $\Phi_\omega(\mathbf{M}) = \inf\limits_{\mathbf{M}=\mathbf{A}+\mathbf{B}} \left\{ \lVert \mathbf{A} \rVert_1 + \omega \lvert\lvert\lvert \mathbf{B} \rvert\rvert\rvert_{\text{nuc}} \right\}$	Weighted max. norm $\Phi^*(\mathbf{N}) = \max \left\{ \lVert \mathbf{N} \rVert_{\max}, \omega^{-1} \lvert\lvert\lvert \mathbf{N} \rvert\rvert\rvert_2 \right\}$

Table 9.1 *Primal and dual pairs of regularizers in various cases. See Exercises 9.4 and 9.5 for verification of some of these correspondences.*

Table 9.1 gives some examples of various dual norm pairs.

Our choice of regularization parameter is specified in terms of the random vector $\nabla \mathcal{L}_n(\theta^*)$ —the gradient of the empirical cost evaluated at θ^*, also referred to as the *score function*. Under mild regularity conditions, we have $\mathbb{E}[\nabla \mathcal{L}_n(\theta^*))] = \nabla \bar{\mathcal{L}}(\theta^*)$. Consequently, when the target parameter θ^* lies in the interior of the parameter space Ω, by the optimality conditions for the minimization (9.2), the random vector $\nabla \mathcal{L}_n(\theta^*)$ has zero mean. Under ideal circumstances, we expect that the score function will not be too large, and we measure its fluctuations in terms of the dual norm, thereby defining the "good event"

$$\mathbb{G}(\lambda_n) := \left\{ \Phi^*(\nabla \mathcal{L}_n(\theta^*)) \leq \frac{\lambda_n}{2} \right\}. \tag{9.28}$$

With this set-up, we are now ready for the statement of the main technical result of this section. The reader should recall the definition of the subspace projection operator (9.20).

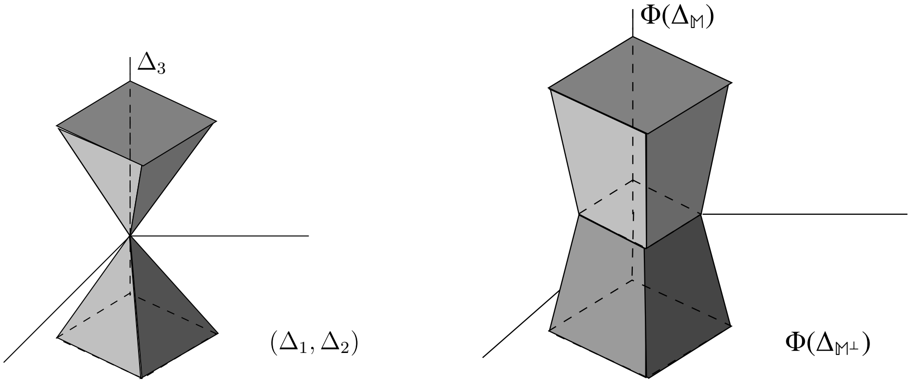
Proposition 9.13 *Let $\mathcal{L}_n \colon \Omega \to \mathbb{R}$ be a convex function, let the regularizer $\Phi \colon \Omega \to [0, \infty)$ be a norm, and consider a subspace pair $(\mathbb{M}, \overline{\mathbb{M}}^\perp)$ over which Φ is decomposable. Then conditioned on the event $\mathbb{G}(\lambda_n)$, the error $\widehat{\Delta} = \widehat{\theta} - \theta^*$ belongs to the set*

$$\mathbb{C}_{\theta^*}(\mathbb{M}, \overline{\mathbb{M}}^\perp) := \{\Delta \in \Omega \mid \Phi(\Delta_{\overline{\mathbb{M}}^\perp}) \leq 3\Phi(\Delta_{\overline{\mathbb{M}}}) + 4\Phi(\theta^*_{\mathbb{M}^\perp})\}. \tag{9.29}$$

When the subspaces $(\mathbb{M}, \overline{\mathbb{M}}^\perp)$ and parameter θ^* are clear from the context, we adopt the shorthand notation $\mathbb{C}$. Figure 9.7 provides an illustration of the geometric structure of the set $\mathbb{C}$. To understand its significance, let us consider the special case when $\theta^* \in \mathbb{M}$, so that $\theta^*_{\mathbb{M}^\perp} = 0$. In this case, membership of $\widehat{\Delta}$ in $\mathbb{C}$ implies that $\Phi(\widehat{\Delta}_{\bar{\mathbb{M}}^\perp}) \leq 3\Phi(\widehat{\Delta}_{\bar{\mathbb{M}}})$, and hence that

$$\Phi(\widehat{\Delta}) = \Phi(\widehat{\Delta}_{\bar{\mathbb{M}}} + \widehat{\Delta}_{\bar{\mathbb{M}}^\perp}) \leq \Phi(\widehat{\Delta}_{\bar{\mathbb{M}}}) + \Phi(\widehat{\Delta}_{\bar{\mathbb{M}}^\perp}) \leq 4\Phi(\widehat{\Delta}_{\bar{\mathbb{M}}}). \tag{9.30}$$

Consequently, when measured in the norm defined by the regularizer, the vector $\widehat{\Delta}$ is only a constant factor larger than the projected quantity $\widehat{\Delta}_{\bar{\mathbb{M}}}$. Whenever the subspace $\bar{\mathbb{M}}$ is relatively small, this inequality provides significant control on $\widehat{\Delta}$.


Δ_3
(Δ_1, Δ_2)
$\Phi(\Delta_{\mathbb{M}})$
$\Phi(\Delta_{\mathbb{M}^\perp})$

(a) (b)

Figure 9.7 Illustration of the set $\mathbb{C}_{\theta^*}(\mathbb{M}, \bar{\mathbb{M}}^\perp)$ in the special case $\Delta = (\Delta_1, \Delta_2, \Delta_3) \in \mathbb{R}^3$ and regularizer $\Phi(\Delta) = \|\Delta\|_1$, relevant for sparse vectors (Example 9.1). This picture shows the case $S = \{3\}$, so that the model subspace is $\mathbb{M}(S) = \{\Delta \in \mathbb{R}^3 \mid \Delta_1 = \Delta_2 = 0\}$, and its orthogonal complement is given by $\mathbb{M}^\perp(S) = \{\Delta \in \mathbb{R}^3 \mid \Delta_3 = 0\}$. (a) In the special case when $\theta^*_1 = \theta^*_2 = 0$, so that $\theta^* \in \mathbb{M}$, the set $\mathbb{C}(\mathbb{M}, \mathbb{M}^\perp)$ is a cone, with no dependence on θ^*. (b) When θ^* does not belong to $\mathbb{M}$, the set $\mathbb{C}(\mathbb{M}, \mathbb{M}^\perp)$ is enlarged in the coordinates (Δ_1, Δ_2) that span $\mathbb{M}^\perp$. It is no longer a cone, but is still a star-shaped set.

We now turn to the proof of the proposition:

Proof Our argument is based on the function $\mathcal{F} : \Omega \to \mathbb{R}$ given by

$$\mathcal{F}(\Delta) := \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) + \lambda_n\{\Phi(\theta^* + \Delta) - \Phi(\theta^*)\}. \tag{9.31}$$

By construction, we have $\mathcal{F}(0) = 0$, and so the optimality of $\widehat{\theta}$ implies that the error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ must satisfy the condition $\mathcal{F}(\widehat{\Delta}) \leq 0$, corresponding to a basic inequality in this general setting. Our goal is to exploit this fact in order to establish the inclusion (9.29). In order to do so, we require control on the two separate pieces of $\mathcal{F}$, as summarized in the following:

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

Lemma 9.14 (Deviation inequalities) *For any decomposable regularizer and parameters θ^* and Δ, we have*

$$\Phi(\theta^* + \Delta) - \Phi(\theta^*) \geq \Phi(\Delta_{\bar{\mathbb{M}}^\perp}) - \Phi(\Delta_{\bar{\mathbb{M}}}) - 2\Phi(\theta^*_{\mathbb{M}^\perp}). \tag{9.32}$$

Moreover, for any convex function $\mathcal{L}_n$, conditioned on the event $\mathbb{G}(\lambda_n)$, we have

$$\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) \geq -\frac{\lambda_n}{2}\left[\Phi\left(\Delta_{\bar{\mathbb{M}}}\right) + \Phi\left(\Delta_{\bar{\mathbb{M}}^\perp}\right)\right]. \tag{9.33}$$

Given this lemma, the claim of Proposition 9.13 follows immediately. Indeed, combining the two lower bounds (9.32) and (9.33), we obtain

$$\begin{aligned} 0 \geq \mathcal{F}(\widehat{\Delta}) &\geq \lambda_n \left\{\Phi(\Delta_{\bar{\mathbb{M}}^\perp}) - \Phi(\Delta_{\bar{\mathbb{M}}}) - 2\Phi(\theta^*_{\mathbb{M}^\perp})\right\} - \frac{\lambda_n}{2}\left\{\Phi\left(\Delta_{\bar{\mathbb{M}}}\right) + \Phi\left(\Delta_{\bar{\mathbb{M}}^\perp}\right)\right\} \\ &= \frac{\lambda_n}{2}\left\{\Phi(\Delta_{\bar{\mathbb{M}}^\perp}) - 3\Phi(\Delta_{\bar{\mathbb{M}}}) - 4\Phi(\theta^*_{\mathbb{M}^\perp})\right\}, \end{aligned}$$

from which the claim follows.

Thus, it remains to prove Lemma 9.14, and here we exploit decomposability of the regularizer. Since $\Phi\left(\theta^* + \Delta\right) = \Phi\left(\theta^*_{\mathbb{M}} + \theta^*_{\mathbb{M}^\perp} + \Delta_{\bar{\mathbb{M}}} + \Delta_{\bar{\mathbb{M}}^\perp}\right)$, applying the triangle inequality yields

$$\Phi\left(\theta^* + \Delta\right) \geq \Phi\left(\theta^*_{\mathbb{M}} + \Delta_{\bar{\mathbb{M}}^\perp}\right) - \Phi\left(\theta^*_{\mathbb{M}^\perp} + \Delta_{\bar{\mathbb{M}}}\right) \geq \Phi\left(\theta^*_{\mathbb{M}} + \Delta_{\bar{\mathbb{M}}^\perp}\right) - \Phi\left(\theta^*_{\mathbb{M}^\perp}\right) - \Phi\left(\Delta_{\bar{\mathbb{M}}}\right).$$

By decomposability applied to $\theta^*_{\mathbb{M}}$ and $\Delta_{\bar{\mathbb{M}}^\perp}$, we have $\Phi\left(\theta^*_{\mathbb{M}} + \Delta_{\bar{\mathbb{M}}^\perp}\right) = \Phi\left(\theta^*_{\mathbb{M}}\right) + \Phi\left(\Delta_{\bar{\mathbb{M}}^\perp}\right)$, so that

$$\Phi\left(\theta^* + \Delta\right) \geq \Phi\left(\theta^*_{\mathbb{M}}\right) + \Phi\left(\Delta_{\bar{\mathbb{M}}^\perp}\right) - \Phi\left(\theta^*_{\mathbb{M}^\perp}\right) - \Phi\left(\Delta_{\bar{\mathbb{M}}}\right). \tag{9.34}$$

Similarly, by the triangle inequality, we have $\Phi(\theta^*) \leq \Phi\left(\theta^*_{\mathbb{M}}\right) + \Phi\left(\theta^*_{\mathbb{M}^\perp}\right)$. Combining this inequality with the bound (9.34), we obtain

$$\begin{aligned} \Phi\left(\theta^* + \Delta\right) - \Phi(\theta^*) &\geq \Phi\left(\theta^*_{\mathbb{M}}\right) + \Phi\left(\Delta_{\bar{\mathbb{M}}^\perp}\right) - \Phi\left(\theta^*_{\mathbb{M}^\perp}\right) - \Phi\left(\Delta_{\bar{\mathbb{M}}}\right) - \left\{\Phi\left(\theta^*_{\mathbb{M}}\right) + \Phi\left(\theta^*_{\mathbb{M}^\perp}\right)\right\} \\ &= \Phi\left(\Delta_{\bar{\mathbb{M}}^\perp}\right) - \Phi\left(\Delta_{\bar{\mathbb{M}}}\right) - 2\Phi\left(\theta^*_{\mathbb{M}^\perp}\right), \end{aligned}$$

which yields the claim (9.32).

Turning to the cost difference, using the convexity of the cost function $\mathcal{L}_n$, we have

$$\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) \geq \langle \nabla \mathcal{L}_n(\theta^*),\, \Delta\rangle \geq -|\langle \nabla \mathcal{L}_n(\theta^*),\, \Delta\rangle|.$$

Applying the Hölder inequality with the regularizer and its dual (see Exercise 9.7), we have

$$|\langle \nabla \mathcal{L}_n(\theta^*),\, \Delta\rangle| \leq \Phi^*(\nabla \mathcal{L}_n(\theta^*))\, \Phi(\Delta) \leq \frac{\lambda_n}{2}\left[\Phi\left(\Delta_{\bar{\mathbb{M}}}\right) + \Phi\left(\Delta_{\bar{\mathbb{M}}^\perp}\right)\right],$$

where the final step uses the triangle inequality, and the assumed bound $\lambda_n \geq 2\Phi^*(\nabla \mathcal{L}_n(\theta^*))$. Putting together the pieces yields the claimed bound (9.33). This completes the proof of Lemma 9.14, and hence the proof of the proposition. □

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

9.3 Restricted curvature conditions

We now turn to the second component of a general framework, which concerns the curvature of the cost function. Before discussing the general high-dimensional setting, let us recall the classical role of curvature in maximum likelihood estimation, where it enters via the Fisher information matrix. Under i.i.d. sampling, the principle of maximum likelihood is equivalent to minimizing the cost function

$$\mathcal{L}_n(\theta) := -\frac{1}{n}\sum_{i=1}^{n}\log\mathbb{P}_\theta(z_i). \tag{9.35}$$

The Hessian of this cost function $\nabla^2\mathcal{L}_n(\theta)$ is the sample version of the Fisher information matrix; as the sample size n increases to infinity with d fixed, it converges in a pointwise sense to the population *Fisher information* $\nabla^2\overline{\mathcal{L}}(\theta)$. Recall that the population cost function $\overline{\mathcal{L}}$ was defined previously in equation (9.1). The Fisher information matrix evaluated at θ^* provides a lower bound on the accuracy of any statistical estimator via the Cramér–Rao bound. As a second derivative, the Fisher information matrix $\nabla^2\overline{\mathcal{L}}(\theta^*)$ captures the curvature of the cost function around the point θ^*.

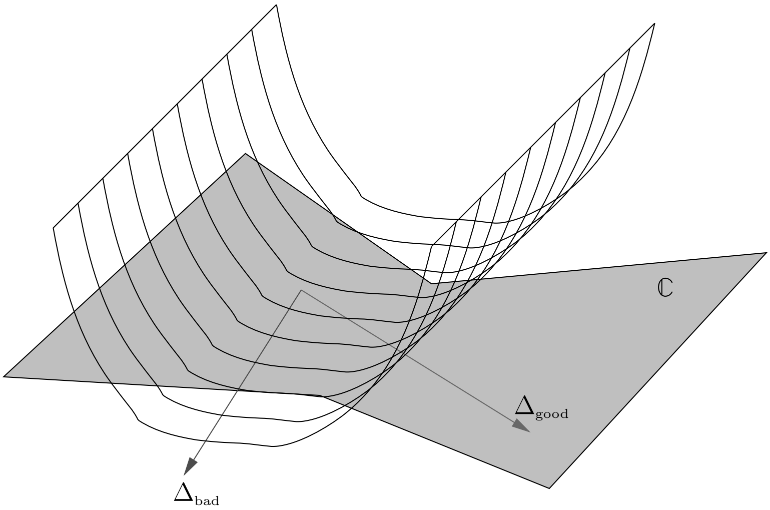


Figure 9.8 Illustration of the cost function $\theta \mapsto \mathcal{L}_n(\theta; Z_1^n)$. In the high-dimensional setting ($d > n$), although it may be curved in certain directions (e.g., Δ_{good}), there are $d - n$ directions in which it is flat up to second order (e.g., Δ_{bad}).

In the high-dimensional setting, the story becomes a little more complicated. In particular, whenever $n < d$, then the sample Fisher information matrix $\nabla^2\mathcal{L}_n(\theta^*)$ is rank-degenerate. Geometrically, this rank degeneracy implies that the cost function takes the form shown in Figure 9.8: while curved upwards in certain directions, there are $d - n$ directions in which it is flat up to second order. Consequently, the high-dimensional setting precludes any type of uniform lower bound on the curvature, and we can only hope to obtain some form of *restricted curvature*. There are several ways in which to develop such notions, and we describe two in the sections to follow, the first based on lower bounding the error in the first-order

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

Taylor-series expansion, and the second by directly lower bounding the curvature of the gradient mapping.

9.3.1 Restricted strong convexity

We begin by describing the notion of restricted strong convexity, which is defined by the Taylor-series expansion. Given any differentiable cost function, we can use the gradient to form the first-order Taylor approximation, which then defines the *first-order Taylor-series error*

$$\mathcal{E}_n(\Delta) := \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \langle \nabla \mathcal{L}_n(\theta^*),\, \Delta \rangle\,. \tag{9.36}$$

Whenever the function $\theta \mapsto \mathcal{L}_n(\theta)$ is convex, this error term is always guaranteed to be non-negative.[2] Strong convexity requires that this lower bound holds with a quadratic slack: in particular, for a given norm $\|\cdot\|$, the cost function is locally *κ-strongly convex* at θ^* if the first-order Taylor error is lower bounded as

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2 \tag{9.37}$$

for all Δ in a neighborhood of the origin. As previously discussed, this notion of strong convexity cannot hold for a generic high-dimensional problem. But for decomposable regularizers, we have seen (Proposition 9.13) that the error vector must belong to a very special set, and we use this fact to define the notion of restricted strong convexity.

Definition 9.15 For a given norm $\|\cdot\|$ and regularizer $\Phi(\cdot)$, the cost function satisfies a *restricted strong convexity* (RSC) condition with radius $R > 0$, curvature $\kappa > 0$ and tolerance τ_n^2 if

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\,\|\Delta\|^2 - \tau_n^2\,\Phi^2(\Delta) \qquad \text{for all } \Delta \in \mathbb{B}(R). \tag{9.38}$$

To clarify a few aspects of this definition, the set $\mathbb{B}(R)$ is the unit ball defined by the given norm $\|\cdot\|$. In our applications of RSC, the norm $\|\cdot\|$ will be derived from an inner product on the space Ω. Standard cases include the usual Euclidean norm on $\mathbb{R}^d$, and the Frobenius norm on the matrix space $\mathbb{R}^{d_1 \times d_2}$. Various types of weighted quadratic norms also fall within this general class.

Note that, if we set the tolerance term $\tau_n^2 = 0$, then the RSC condition (9.38) is equivalent to asserting that $\mathcal{L}_n$ is locally strongly convex in a neighborhood of θ^* with coefficient κ. As previously discussed, such a strong convexity condition cannot hold in the high-dimensional setting. However, given our goal of proving error bounds on M-estimators, we are not interested in all directions, but rather only the directions in which the error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ can lie. For decomposable regularizers, Proposition 9.13 guarantees that the error vector must lie in the very special "cone-like" sets $\mathbb{C}_{\theta^*}(\mathbb{M}, \overline{\mathbb{M}}^\perp)$. Even with a strictly positive tolerance $\tau_n^2 > 0$, an RSC condition of the form (9.38) can be used to guarantee a lower curvature over

[2] Indeed, for differentiable functions, this property may be viewed as an equivalent definition of convexity.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

this restricted set, as long as the sample size is sufficiently large. We formalize this intuition after considering a few concrete instances of Definition 9.15.

Example 9.16 (Restricted eigenvalues for least-squares cost) In this example, we show how the restricted eigenvalue conditions (see Definition 7.12 in Chapter 7) correspond to a special case of restricted strong convexity. For the least-squares objective $\mathcal{L}_n(\theta) = \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2$, an easy calculation yields that the first-order Taylor error is given by $\mathcal{E}_n(\Delta) = \frac{\|\mathbf{X}\Delta\|_2^2}{2n}$. A restricted strong convexity condition with the ℓ_1-norm then takes the form

$$\frac{\|\mathbf{X}\Delta\|_2^2}{2n} \geq \frac{\kappa}{2}\|\Delta\|_2^2 - \tau_n^2\|\Delta\|_1^2 \qquad \text{for all } \Delta \in \mathbb{R}^d. \tag{9.39}$$

For various types of sub-Gaussian matrices, bounds of this form hold with high probability for the choice $\tau_n^2 \asymp \frac{\log d}{n}$. Theorem 7.16 in Chapter 7 provides one instance of such a result.

As a side remark, this example shows that the least-squares objective is special in two ways: the first-order Taylor error is independent of θ^* and, moreover, it is a positively homogeneous function of degree two—that is, $\mathcal{E}_n(t\Delta) = t^2\mathcal{E}_n(\Delta)$ for all $t \in \mathbb{R}$. The former property implies that we need not be concerned about uniformity in θ^*, whereas the latter implies that it is not necessary to localize Δ to a ball $\mathbb{B}(R)$. ♣

Later in Section 9.8, we provide more general results, showing that a broader class of cost functions satisfy a restricted strong convexity condition of the type (9.39). Let us consider one example here:

Example 9.17 (RSC for generalized linear models) Recall the family of generalized linear models from Example 9.2, and the cost function (9.7) defined by the negative log-likelihood. Suppose that we draw n i.i.d. samples, in which the covariates $\{x_i\}_{i=1}^n$ are drawn from a zero-mean sub-Gaussian distribution with non-degenerate covariance matrix $\mathbf{\Sigma}$. As a consequence of a result to follow (Theorem 9.36), the Taylor-series error of various GLM log-likelihoods satisfies a lower bound of the form

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_2^2 - c_1\frac{\log d}{n}\|\Delta\|_1^2 \qquad \text{for all } \|\Delta\|_2 \leq 1 \tag{9.40}$$

with probability greater than $1 - c_2\exp(-c_3 n)$.

Theorem 9.36 actually provides a more general guarantee in terms of the quantity

$$\mu_n(\Phi^*) := \mathbb{E}_{x,\varepsilon}\left[\Phi^*\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i\right)\right], \tag{9.41}$$

where Φ^* denotes the dual norm, and $\{\varepsilon_i\}_{i=1}^n$ is a sequence of i.i.d. Rademacher variables. With this notation, we have

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_2^2 - c_1\,\mu_n^2(\Phi^*)\,\Phi^2(\Delta) \qquad \text{for all } \|\Delta\|_2 \leq 1 \tag{9.42}$$

with probability greater than $1 - c_2\exp(-c_3 n)$. This result is a generalization of our previous bound (9.40), since $\mu_n(\Phi^*) \precsim \sqrt{\frac{\log d}{n}}$ in the case of ℓ_1-regularization.

In Exercise 9.8, we bound the quantity (9.41) for various norms. For group Lasso with group set $\mathcal{G}$ and maximum group size m, we show that

$$\mu_n(\Phi^*) \precsim \sqrt{\frac{m}{n}} + \sqrt{\frac{\log |\mathcal{G}|}{n}}, \tag{9.43a}$$

whereas for the nuclear norm for $d_1 \times d_2$ matrices, we show that

$$\mu_n(\Phi^*) \precsim \sqrt{\frac{d_1}{n}} + \sqrt{\frac{d_2}{n}}. \tag{9.43b}$$

We also show how these results, in conjunction with the lower bound (9.42), imply suitable forms of restricted convexity as long as the sample size is sufficiently large. ♣

We conclude this section with the definition of one last geometric parameter that plays an important role. As we have just seen, in the context of ℓ_1-regularization and the RE condition, the cone constraint is very useful; in particular, it implies that $\|\Delta\|_1 \leq 4\sqrt{s}\|\Delta\|_2$, a bound used repeatedly in Chapter 7. Returning to the general setting, we need to study how to translate between $\Phi(\Delta_{\mathbb{M}})$ and $\|\Delta_{\mathbb{M}}\|$ for an arbitrary decomposable regularizer and error norm.

Definition 9.18 (Subspace Lipschitz constant) For any subspace $\mathbb{S}$ of $\mathbb{R}^d$, the *subspace Lipschitz constant* with respect to the pair $(\Phi, \|\cdot\|)$ is given by

$$\Psi(\mathbb{S}) := \sup_{u \in \mathbb{S}\setminus\{0\}} \frac{\Phi(u)}{\|u\|}. \tag{9.44}$$

To clarify our terminology, this quantity is the Lipschitz constant of the regularizer with respect to the error norm, but as restricted to the subspace $\mathbb{S}$. It corresponds to the worst-case price of translating between the Φ- and $\|\cdot\|$-norms for any vector in $\mathbb{S}$.

To illustrate its use, let us consider it in the special case when $\theta^* \in \mathbb{M}$. Then for any $\Delta \in \mathbb{C}_{\theta^*}(\mathbb{M}, \bar{\mathbb{M}}^\perp)$, we have

$$\Phi(\Delta) \overset{\text{(i)}}{\leq} \Phi(\Delta_{\bar{\mathbb{M}}}) + \Phi(\Delta_{\bar{\mathbb{M}}^\perp}) \overset{\text{(ii)}}{\leq} 4\Phi(\Delta_{\bar{\mathbb{M}}}) \overset{\text{(iii)}}{\leq} 4\,\Psi(\bar{\mathbb{M}})\|\Delta\|, \tag{9.45}$$

where step (i) follows from the triangle inequality, step (ii) from membership in $\mathbb{C}(\mathbb{M}, \mathbb{M}^\perp)$, and step (iii) from the definition of $\Psi(\bar{\mathbb{M}})$.

As a simple example, if $\mathbb{M}$ is a subspace of s-sparse vectors, then with regularizer $\Phi(u) = \|u\|_1$ and error norm $\|u\| = \|u\|_2$, we have $\Psi(\mathbb{M}) = \sqrt{s}$. In this way, we see that inequality (9.45) is a generalization of the familiar inequality $\|\Delta\|_2 \leq 4\sqrt{s}\|\Delta\|_1$ in the context of sparse vectors. The subspace Lipschitz constant appears explicitly in the main results, and also arises in establishing restricted strong convexity.

9.4 Some general theorems

Thus far, we have discussed the notion of decomposable regularizers, and some related notions of restricted curvature for the cost function. In this section, we state and prove some

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

results on the estimation error, namely, the quantity $\widehat{\theta} - \theta^*$, where $\widehat{\theta}$ denotes any optimum of the regularized M-estimator (9.3).

9.4.1 Guarantees under restricted strong convexity

We begin by stating and proving a general result that holds under the restricted strong convexity condition given in Section 9.3.1. Let us summarize the assumptions that we impose throughout this section:

(A1) The cost function is convex, and satisfies the local RSC condition (9.38) with curvature κ, radius R and tolerance τ_n^2 with respect to an inner-product induced norm $\|\cdot\|$.

(A2) There is a pair of subspaces $\mathbb{M} \subseteq \overline{\mathbb{M}}$ such that the regularizer decomposes over $(\mathbb{M}, \overline{\mathbb{M}}^\perp)$.

We state the result as a deterministic claim, but conditioned on the "good" event

$$\mathbb{G}(\lambda_n) := \left\{ \Phi^*(\nabla \mathcal{L}_n(\theta^*)) \leq \frac{\lambda_n}{2} \right\}. \tag{9.46}$$

Our bound involves the quantity

$$\varepsilon_n^2(\overline{\mathbb{M}}, \mathbb{M}^\perp) := \underbrace{9\,\frac{\lambda_n^2}{\kappa^2}\,\Psi^2(\overline{\mathbb{M}})}_{\text{estimation error}} + \underbrace{\frac{8}{\kappa}\,\left\{\lambda_n \Phi(\theta^*_{\mathbb{M}^\perp}) + 16\tau_n^2 \Phi^2(\theta^*_{\mathbb{M}^\perp})\right\}}_{\text{approximation error}}, \tag{9.47}$$

which depends on the choice of our subspace pair $(\overline{\mathbb{M}}, \mathbb{M}^\perp)$.

Theorem 9.19 (Bounds for general models) *Under conditions (A1) and (A2), consider the regularized M-estimator* (9.3) *conditioned on the event $\mathbb{G}(\lambda_n)$,*

(a) *Any optimal solution satisfies the bound*

$$\Phi(\widehat{\theta} - \theta^*) \leq 4\,\left\{\Psi(\overline{\mathbb{M}})\,\|\widehat{\theta} - \theta^*\| + \Phi(\theta^*_{\mathbb{M}^\perp})\right\}. \tag{9.48a}$$

(b) *For any subspace pair $(\overline{\mathbb{M}}, \mathbb{M}^\perp)$ such that $\tau_n^2 \Psi^2(\overline{\mathbb{M}}) \leq \frac{\kappa}{64}$ and $\varepsilon_n(\overline{\mathbb{M}}, \mathbb{M}^\perp) \leq R$, we have*

$$\|\widehat{\theta} - \theta^*\|^2 \leq \varepsilon_n^2(\overline{\mathbb{M}}, \mathbb{M}^\perp). \tag{9.48b}$$

It should be noted that Theorem 9.19 is actually a deterministic result. Probabilistic conditions enter in certifying that the RSC condition holds with high probability (see Section 9.8), and in verifying that, for a concrete choice of regularization parameter, the *dual norm bound* $\lambda_n \geq 2\Phi^*(\nabla \mathcal{L}_n(\theta^*))$ defining the event $\mathbb{G}(\lambda_n)$ holds with high probability. The dual norm bound cannot be explicitly verified, since it presumes knowledge of θ^*, but it suffices to give choices of λ_n for which it holds with high probability. We illustrate such choices in various examples to follow.

Equations (9.48a) and (9.48b) actually specify a family of upper bounds, one for each subspace pair $(\overline{\mathbb{M}}, \overline{\mathbb{M}}^\perp)$ over which the regularizer Φ decomposes. The optimal choice of these subspaces serves to trade off the estimation and approximation error terms in the bound. The upper bound (9.48b) corresponds to an *oracle inequality,* since it applies to any parameter θ^*, and gives a family of upper bounds involving two sources of error. The term labeled "estimation error" represents the statistical cost of estimating a parameter belong to the subspace $\mathbb{M} \subseteq \bar{\mathbb{M}}$; naturally, it increases as $\mathbb{M}$ grows. The second quantity represents "approximation error" incurred by estimating only within the subspace $\mathbb{M}$, and it shrinks as $\mathbb{M}$ is increased. Thus, the optimal bound is obtained by choosing the model subspace to balance these two types of error. We illustrate such choices in various examples to follow.

In the special case that the target parameter θ^* is contained within a subspace $\mathbb{M}$, Theorem 9.19 has the following corollary:

Corollary 9.20 *Suppose that, in addition to the conditions of Theorem 9.19, the optimal parameter θ^* belongs to $\mathbb{M}$. Then any optimal solution $\widehat{\theta}$ to the optimization problem* (9.3) *satisfies the bounds*

$$\Phi(\widehat{\theta} - \theta^*) \leq 6\frac{\lambda_n}{\kappa}\Psi^2(\bar{\mathbb{M}}), \tag{9.49a}$$

$$\|\widehat{\theta} - \theta^*\|^2 \leq 9\,\frac{\lambda_n^2}{\kappa^2}\Psi^2(\bar{\mathbb{M}}). \tag{9.49b}$$

This corollary can be applied directly to obtain concrete estimation error bounds for many problems, as we illustrate in the sequel.

We now turn to the proof of Theorem 9.19.

Proof We begin by proving part (a). Letting $\widehat{\Delta} = \widehat{\theta} - \theta^*$ be the error, by the triangle inequality, we have

$$\begin{aligned}
\Phi(\widehat{\Delta}) &\leq \Phi(\widehat{\Delta}_{\bar{\mathbb{M}}}) + \Phi(\widehat{\Delta}_{\bar{\mathbb{M}}^\perp}) \\
&\overset{\text{(i)}}{\leq} \Phi(\widehat{\Delta}_{\bar{\mathbb{M}}}) + \Big\{3\Phi(\widehat{\Delta}_{\bar{\mathbb{M}}}) + 4\Phi(\theta^*_{\mathbb{M}^\perp})\Big\} \\
&\overset{\text{(ii)}}{\leq} 4\Big\{\Psi(\bar{\mathbb{M}})\,\|\widehat{\theta} - \theta^*\| + \Phi(\theta^*_{\mathbb{M}^\perp})\Big\},
\end{aligned}$$

where inequality (i) follows from Proposition 9.13 under event $\mathbb{G}(\lambda_n)$ and inequality (ii) follows from the definition of the optimal subspace constant.

Turning to the proof of part (b), in order to simplify notation, we adopt the shorthand $\mathbb{C}$ for the set $\mathbb{C}_{\theta^*}(\mathbb{M}, \bar{\mathbb{M}}^\perp)$. Letting $\delta \in (0, R]$ be a given error radius to be chosen, the following lemma shows that it suffices to control the sign of the function $\mathcal{F}$ from equation (9.31) over the set $\mathbb{K}(\delta) := \mathbb{C} \cap \{\|\Delta\| = \delta\}$.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

Lemma 9.21 *If $\mathcal{F}(\Delta) > 0$ for all vectors $\Delta \in \mathbb{K}(\delta)$, then $\|\widehat{\Delta}\| \le \delta$.*

Proof We prove the contrapositive statement: in particular, we show that if for some optimal solution $\widehat{\theta}$, the associated error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ satisfies the inequality $\|\widehat{\Delta}\| > \delta$, then there must be some vector $\widetilde{\Delta} \in \mathbb{K}(\delta)$ such that $\mathcal{F}(\widetilde{\Delta}) \le 0$. If $\|\widehat{\Delta}\| > \delta$, then since $\mathbb{C}$ is star-shaped around the origin (see the Appendix, Section 9.9), the line joining $\widehat{\Delta}$ to 0 must intersect the set $\mathbb{K}(\delta)$ at some intermediate point of the form $t^*\widehat{\Delta}$ for some $t^* \in [0, 1]$. See Figure 9.9 for an illustration.

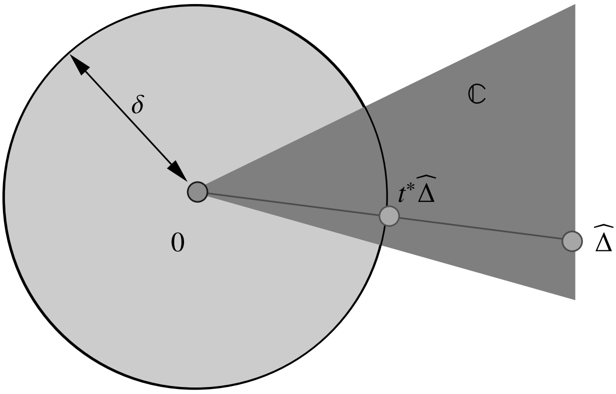


Figure 9.9 Geometry of the proof of Lemma 9.21. When $\|\widehat{\Delta}\| > \delta$ and the set $\mathbb{C}$ is star-shaped around the origin, any line joining $\widehat{\Delta}$ and the origin 0 must intersect the set $\mathbb{K}(\delta) = \{\|\Delta\| = \delta\} \cap \mathbb{C}$ at some intermediate point of the form $t^*\widehat{\Delta}$ for some $t^* \in [0, 1]$.

Since the cost function $\mathcal{L}_n$ and regularizer Φ are convex, the function $\mathcal{F}$ is also convex for any non-negative choice of the regularization parameter. Given the convexity of $\mathcal{F}$, we can apply Jensen's inequality so as to obtain

$$\mathcal{F}(t^*\widehat{\Delta}) = \mathcal{F}(t^*\widehat{\Delta} + (1-t^*)\,0) \le t^*\,\mathcal{F}(\widehat{\Delta}) + (1-t^*)\mathcal{F}(0) \overset{\text{(i)}}{=} t^*\mathcal{F}(\widehat{\Delta}),$$

where equality (i) uses the fact that $\mathcal{F}(0) = 0$ by construction. But since $\widehat{\Delta}$ is optimal, we must have $\mathcal{F}(\widehat{\Delta}) \le 0$, and hence $\mathcal{F}(t^*\Delta) \le 0$ as well. Thus, we have constructed a vector $\widetilde{\Delta} = t^*\Delta$ with the claimed properties, thereby establishing the claim in the lemma. □

We now return to the proof of Theorem 9.19. Fix some radius $\delta \in (0, R]$, whose value will be specified later in the proof (see equation (9.53)). On the basis of Lemma 9.21, the proof of Theorem 9.19 will be complete if we can establish a lower bound on the function value

$\mathcal{F}(\Delta)$ for all vectors $\Delta \in \mathbb{K}(\delta)$. For an arbitrary vector $\Delta \in \mathbb{K}(\delta)$, we have

$$\begin{aligned}\mathcal{F}(\Delta) &= \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) + \lambda_n\{\Phi(\theta^* + \Delta) - \Phi(\theta^*)\} \\ &\overset{\text{(i)}}{\geq} \langle \nabla\mathcal{L}_n(\theta^*),\, \Delta\rangle + \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2\Phi^2(\Delta) + \lambda_n\{\Phi(\theta^* + \Delta) - \Phi(\theta^*)\} \qquad (9.50) \\ &\overset{\text{(ii)}}{\geq} \langle \nabla\mathcal{L}_n(\theta^*),\, \Delta\rangle + \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2\Phi^2(\Delta) + \lambda_n\{\Phi(\Delta_{\bar{\mathbb{M}}^\perp}) - \Phi(\Delta_{\bar{\mathbb{M}}}) - 2\Phi(\theta^*_{\mathbb{M}^\perp})\},\end{aligned}$$

where inequality (i) follows from the RSC condition, and inequality (ii) follows from the bound (9.32).

By applying Hölder's inequality with the regularizer Φ and its dual Φ^*, we find that

$$|\langle \nabla\mathcal{L}_n(\theta^*),\, \Delta\rangle| \leq \Phi^*(\nabla\mathcal{L}_n(\theta^*))\; \Phi(\Delta).$$

Under the event $\mathbb{G}(\lambda_n)$, the regularization parameter is lower bounded as $\lambda_n \geq 2\Phi^*(\nabla\mathcal{L}_n(\theta^*))$, which implies that $|\langle \nabla\mathcal{L}_n(\theta^*),\, \Delta\rangle| \leq \frac{\lambda_n}{2}\Phi(\Delta)$. Consequently, we have

$$\mathcal{F}(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2\Phi^2(\Delta) + \lambda_n\Big\{\Phi(\Delta_{\bar{\mathbb{M}}^\perp}) - \Phi(\Delta_{\bar{\mathbb{M}}}) - 2\Phi(\theta^*_{\mathbb{M}^\perp})\Big\} - \frac{\lambda_n}{2}\Phi(\Delta).$$

The triangle inequality implies that

$$\Phi(\Delta) = \Phi(\Delta_{\bar{\mathbb{M}}^\perp} + \Delta_{\bar{\mathbb{M}}}) \leq \Phi(\Delta_{\bar{\mathbb{M}}^\perp}) + \Phi(\Delta_{\bar{\mathbb{M}}}),$$

and hence, following some algebra, we find that

$$\begin{aligned}\mathcal{F}(\Delta) &\geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2\Phi^2(\Delta) + \lambda_n\Big\{\frac{1}{2}\Phi(\Delta_{\bar{\mathbb{M}}^\perp}) - \frac{3}{2}\Phi(\Delta_{\bar{\mathbb{M}}}) - 2\Phi(\theta^*_{\mathbb{M}^\perp})\Big\} \\ &\geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2\Phi^2(\Delta) - \frac{\lambda_n}{2}\Big\{3\Phi(\Delta_{\bar{\mathbb{M}}}) + 4\Phi(\theta^*_{\mathbb{M}^\perp})\Big\}. \qquad (9.51)\end{aligned}$$

Now definition (9.44) of the subspace Lipschitz constant implies that $\Phi(\Delta_{\bar{\mathbb{M}}}) \leq \Psi(\bar{\mathbb{M}})\|\Delta_{\bar{\mathbb{M}}}\|$. Since the projection $\Delta \mapsto \Delta_{\bar{\mathbb{M}}}$ is defined in terms of the norm $\|\cdot\|$, it is non-expansive. Since $0 \in \bar{\mathbb{M}}$, we have

$$\|\Delta_{\bar{\mathbb{M}}}\| = \|\Pi_{\bar{\mathbb{M}}}(\Delta) - \Pi_{\bar{\mathbb{M}}}(0)\| \overset{\text{(i)}}{\leq} \|\Delta - 0\| = \|\Delta\|,$$

where inequality (i) uses non-expansiveness of the projection. Combining with the earlier bound, we conclude that $\Phi(\Delta_{\bar{\mathbb{M}}}) \leq \Psi(\bar{\mathbb{M}})\|\Delta\|$.

Similarly, for any $\Delta \in \mathbb{C}$, we have

$$\begin{aligned}\Phi^2(\Delta) \;&\leq \Big\{4\Phi(\Delta_{\bar{\mathbb{M}}}) + 4\Phi(\theta^*_{\mathbb{M}^\perp})\Big\}^2 \leq 32\Phi^2(\Delta_{\bar{\mathbb{M}}}) + 32\Phi^2(\theta^*_{\mathbb{M}^\perp}) \\ &\leq 32\Psi^2(\bar{\mathbb{M}})\,\|\Delta\|^2 + 32\Phi^2(\theta^*_{\mathbb{M}^\perp}). \qquad (9.52)\end{aligned}$$

Substituting into the lower bound (9.51), we obtain the inequality

$$\begin{aligned}\mathcal{F}(\Delta) &\geq \Big\{\frac{\kappa}{2} - 32\tau_n^2\Psi^2(\bar{\mathbb{M}})\Big\}\,\|\Delta\|^2 - 32\tau_n^2\Phi^2(\theta^*_{\mathbb{M}^\perp}) - \frac{\lambda_n}{2}\Big\{3\Psi(\bar{\mathbb{M}})\,\|\Delta\| + 4\Phi(\theta^*_{\mathbb{M}^\perp})\Big\} \\ &\overset{\text{(ii)}}{\geq} \frac{\kappa}{4}\|\Delta\|^2 - \frac{3\lambda_n}{2}\Psi(\bar{\mathbb{M}})\,\|\Delta\| - 32\tau_n^2\Phi^2(\theta^*_{\mathbb{M}^\perp}) - 2\lambda_n\Phi(\theta^*_{\mathbb{M}^\perp}),\end{aligned}$$

where step (ii) uses the assumed bound $\tau_n^2\Psi^2(\bar{\mathbb{M}}) < \frac{\kappa}{64}$.

The right-hand side of this inequality is a strictly positive definite quadratic form in $\|\Delta\|$,

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

and so will be positive for $\|\Delta\|$ sufficiently large. In particular, some algebra shows that this is the case as long as

$$\|\Delta\|^2 \geq \varepsilon_n^2(\overline{\mathbb{M}}, \mathbb{M}^\perp) := 9\, \frac{\lambda_n^2}{\kappa^2}\, \Psi^2(\overline{\mathbb{M}}) + \frac{8}{\kappa}\Big\{\lambda_n \Phi(\theta^*_{\mathbb{M}^\perp}) + 16\tau_n^2 \Phi^2(\theta^*_{\mathbb{M}^\perp})\Big\}. \tag{9.53}$$

This argument is valid as long as $\varepsilon_n \leq R$, as assumed in the statement. □

9.4.2 Bounds under Φ^*-curvature

We now turn to an alternative form of restricted curvature, one which involves a lower bound on the gradient of the cost function. In order to motivate the definition to follow, note that an alternative way of characterizing strong convexity of a differentiable cost function is via the behavior of its gradient. More precisely, a differentiable function $\mathcal{L}_n$ is locally κ-strongly convex at θ^*, in the sense of the earlier definition (9.37), if and only if

$$\langle \nabla \mathcal{L}_n(\theta^* + \Delta)) - \nabla \mathcal{L}_n(\theta^*),\, \Delta \rangle \geq \kappa \|\Delta\|^2 \tag{9.54}$$

for all Δ in some ball around zero. See Exercise 9.9 for verification of the equivalence between the property (9.54) and the earlier definition (9.37). When the underlying norm $\|\cdot\|$ is the ℓ_2-norm, then the condition (9.54), combined with the Cauchy–Schwarz inequality, implies that

$$\|\nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \mathcal{L}_n(\theta^*)\|_2 \geq \kappa \|\Delta\|_2.$$

This implication suggests that it could be useful to consider alternative notions of curvature based on different choices of the norm. Here we consider such a notion based on the dual norm Φ^*:

Definition 9.22 The cost function satisfies a Φ^**-norm curvature condition* with curvature κ, tolerance τ_n and radius R if

$$\Phi^*\Big(\nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \mathcal{L}_n(\theta^*)\Big) \geq \kappa\, \Phi^*(\Delta) - \tau_n \Phi(\Delta) \tag{9.55}$$

for all $\Delta \in \mathbb{B}_{\Phi^*}(R) := \{\theta \in \Omega \mid \Phi^*(\theta) \leq R\}$.

As with restricted strong convexity, this definition is most easily understood in application to the classical case of least-squares cost and ℓ_1-regularization:

Example 9.23 (Restricted curvature for least-squares cost) For the least-squares cost function, we have $\nabla \mathcal{L}_n(\theta) = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\theta - \theta^*) = \widehat{\boldsymbol{\Sigma}}\,(\theta - \theta^*)$, where $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is the sample covariance matrix. For the ℓ_1-norm as the regularizer Φ, the dual norm Φ^* is the ℓ_∞-norm, so that the restricted curvature condition (9.55) is equivalent to the lower bound

$$\Big\|\widehat{\boldsymbol{\Sigma}}\Delta\Big\|_\infty \geq \kappa \|\Delta\|_\infty - \tau_n \|\Delta\|_1 \qquad \text{for all } \Delta \in \mathbb{R}^d. \tag{9.56}$$

In this particular example, localization to the ball $\mathbb{B}_\infty(R)$ is actually unnecessary, since the lower bound is invariant to rescaling of Δ. The bound (9.56) is very closely related to what

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

are known as *ℓ_∞-restricted eigenvalues* of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$. More precisely, such conditions involve lower bounds of the form

$$\left\|\widehat{\boldsymbol{\Sigma}}\Delta\right\|_\infty \geq \kappa' \, \|\Delta\|_\infty \qquad \text{for all } \Delta \in \mathbb{C}(S\,;\alpha), \tag{9.57}$$

where $\mathbb{C}(S\,;\alpha) := \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1\}$, and (κ', α) are given positive constants. In Exercise 9.11, we show that a bound of the form (9.56) implies a form of the ℓ_∞-RE condition (9.57) as long as $n \gtrsim |S|^2 \log d$. Moreover, as we show in Exercise 7.13, such an ℓ_∞-RE condition can be used to derive bounds on the ℓ_∞-error of the Lasso.

Finally, as with ℓ_2-restricted eigenvalue conditions (recall Example 9.16), a lower bound of the form (9.56) holds with high probability with constant κ and tolerance $\tau_n \asymp \sqrt{\frac{\log d}{n}}$ for various types of random design matrices, Exercise 7.14 provides details on one such result. ♣

With this definition in place, we are ready to state the assumptions underlying the main result of this section:

(A1′) The cost satisfies the Φ^*-curvature condition (9.55) with parameters $(\kappa, \tau_n; R)$.

(A2) The regularizer is decomposable with respect to the subspace pair $(\mathbb{M}, \overline{\mathbb{M}}^\perp)$ with $\mathbb{M} \subseteq \overline{\mathbb{M}}$.

Under these conditions, we have the following:

Theorem 9.24 *Given a target parameter $\theta^* \in \mathbb{M}$, consider the regularized M-estimator* (9.3) *under conditions (A1′) and (A2), and suppose that $\tau_n \Psi^2(\overline{\mathbb{M}}) < \frac{\kappa}{32}$. Conditioned on the event $\mathbb{G}(\lambda_n) \cap \{\Phi^*(\widehat{\theta} - \theta^*) \leq R\}$, any optimal solution $\widehat{\theta}$ satisfies the bound*

$$\Phi^*(\widehat{\theta} - \theta^*) \leq 3\,\frac{\lambda_n}{\kappa}. \tag{9.58}$$

Like Theorem 9.19, this claim is deterministic given the stated conditioning. Probabilistic claims enter in certifying that the "good" event $\mathbb{G}(\lambda_n)$ holds with high probability with a specified choice of λ_n. Moreover, except for the special case of least squares, we need to use related results (such as those in Theorem 9.19) to certify that $\Phi^*(\widehat{\theta} - \theta^*) \leq R$, before applying this result.

Proof The proof is relatively straightforward given our development thus far. By standard optimality conditions for a convex program, for any optimum $\widehat{\theta}$, there must exist a subgradient vector $\widehat{z} \in \partial\Phi(\widehat{\theta})$ such that $\nabla\mathcal{L}_n(\widehat{\theta}) + \lambda_n\widehat{z} = 0$. Introducing the error vector $\widehat{\Delta} := \widehat{\theta} - \theta^*$, some algebra yields

$$\nabla\mathcal{L}_n(\theta^* + \widehat{\Delta}) - \nabla\mathcal{L}_n(\theta^*) = -\nabla\mathcal{L}_n(\theta^*) - \lambda_n\widehat{z}.$$

Taking the Φ^*-norm of both sides and applying the triangle inequality yields

$$\Phi^*(\nabla\mathcal{L}_n(\theta^* + \Delta) - \nabla\mathcal{L}_n(\theta^*)) \leq \Phi^*(\nabla\mathcal{L}_n(\theta^*)) + \lambda_n\Phi^*(\widehat{z}).$$

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

On one hand, on the event $\mathbb{G}(\lambda_n)$, we have that $\Phi^*(\nabla \mathcal{L}_n(\theta^*)) \leq \lambda_n/2$, whereas, on the other hand, Exercise 9.6 implies that $\Phi^*(\widetilde{z}) \leq 1$. Putting together the pieces, we find that $\Phi^*(\nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \mathcal{L}_n(\theta^*)) \leq \frac{3\lambda_n}{2}$. Finally, applying the curvature condition (9.55), we obtain

$$\kappa\, \Phi^*(\widehat{\Delta}) \leq \frac{3}{2}\lambda_n + \tau_n \Phi(\widehat{\Delta}). \tag{9.59}$$

It remains to bound $\Phi(\widehat{\Delta})$ in terms of the dual norm $\Phi^*(\widehat{\Delta})$. Since this result is useful in other contexts, we state it as a separate lemma here:

Lemma 9.25 *If $\theta^* \in \mathbb{M}$, then*

$$\Phi(\Delta) \leq 16\Psi^2(\overline{\mathbb{M}})\, \Phi^*(\Delta) \qquad \textit{for any } \Delta \in \mathbb{C}_{\theta^*}(\mathbb{M}, \overline{\mathbb{M}}^\perp). \tag{9.60}$$

Before returning to prove this lemma, we use it to complete the proof of the theorem. On the event $\mathbb{G}(\lambda_n)$, Proposition 9.13 may be applied to guarantee that $\widehat{\Delta} \in \mathbb{C}_{\theta^*}(\mathbb{M}, \overline{\mathbb{M}}^\perp)$. Consequently, the bound (9.60) applies to $\widehat{\Delta}$. Substituting into the earlier bound (9.59), we find that $(\kappa - 16\Psi^2(\mathbb{M})\tau_n)\; \Phi^*(\widehat{\Delta}) \leq \frac{3}{2}\lambda_n$, from which the claim follows by the assumption that $\Psi^2(\mathbb{M})\tau_n \leq \frac{\kappa}{32}$.

We now return to prove Lemma 9.25. From our earlier calculation (9.45), whenever $\theta^* \in \mathbb{M}$ and $\Delta \in \mathbb{C}_{\theta^*}(\mathbb{M}, \overline{\mathbb{M}}^\perp)$, then $\Phi(\Delta) \leq 4\Psi(\overline{\mathbb{M}})\, \|\Delta\|$. Moreover, by Hölder's inequality, we have

$$\|\Delta\|^2 \leq \Phi(\Delta)\, \Phi^*(\Delta) \leq 4\Psi(\overline{\mathbb{M}})\|\Delta\|\, \Phi^*(\Delta),$$

whence $\|\Delta\| \leq 4\Psi(\overline{\mathbb{M}})\Phi^*(\Delta)$. Putting together the pieces, we have

$$\Phi(\Delta) \leq 4\Psi(\overline{\mathbb{M}})\|\Delta\| \leq 16\Psi^2(\overline{\mathbb{M}})\, \Phi^*(\Delta),$$

as claimed. This completes the proof of the lemma, and hence of the theorem. □

Thus far, we have derived two general bounds on the error $\widehat{\theta} - \theta^*$ associated with optima of the M-estimator (9.3). In the remaining sections, we specialize these general results to particular classes of statistical models.

9.5 Bounds for sparse vector regression

We now turn to some consequences of our general theory for the problem of sparse regression. In developing the theory for the full class of generalized linear models, this section provides an alternative and more general complement to our discussion of the sparse linear model in Chapter 7.

9.5.1 Generalized linear models with sparsity

All results in the following two sections are applicable to samples the form $\{(x_i, y_i)\}_{i=1}^n$ where:

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

(G1) The covariates are C-column normalized: $\max_{j=1,\ldots,d} \sqrt{\frac{\sum_{j=1}^d x_{ij}^2}{n}} \leq C$.

(G2) Conditionally on x_i, each response y_i is drawn i.i.d. according to a conditional distribution of the form

$$\mathbb{P}_{\theta^*}(y \mid x) \propto \exp\left\{\frac{y\,\langle x,\, \theta^*\rangle - \psi(\,\langle x,\, \theta^*\rangle\,)}{c(\sigma)}\right\},$$

where the partition function ψ has a bounded second derivative ($\|\psi''\|_\infty \leq B^2$).

We analyze the ℓ_1-regularized version of the GLM log-likelihood estimator, namely

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \Big\{ \underbrace{\frac{1}{n} \sum_{i=1}^n \{\psi(\langle x_i,\, \theta\rangle) - y_i\,\langle x_i,\, \theta\rangle\,\}}_{\mathcal{L}_n(\theta)} + \lambda_n \|\theta\|_1 \Big\}. \tag{9.61}$$

For short, we refer to this M-estimator as the *GLM Lasso*. Note that the usual linear model description $y_i = \langle x_i,\, \theta^*\rangle + w_i$ with $w_i \sim \mathcal{N}(0, \sigma^2)$ falls into this class with $B = 1$, in which the case the estimator (9.61) is equivalent to the ordinary Lasso. It also includes as special cases the problems of logistic regression and multinomial regression, but excludes the case of Poisson regression, due to the boundedness condition (G2).

9.5.2 Bounds under restricted strong convexity

We begin by proving bounds when the Taylor-series error around θ^* associated with the negative log-likelihood (9.61) satisfies the RSC condition

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_2^2 - c_1 \frac{\log d}{n}\|\Delta\|_1^2 \qquad \text{for all } \|\Delta\|_2 \leq 1. \tag{9.62}$$

As discussed in Example 9.17, when the covariates $\{x_i\}_{i=1}^n$ are drawn from a zero-mean sub-Gaussian distribution, a bound of this form holds with high probability for any GLM.

The following result applies to any solution $\widehat{\theta}$ of the GLM Lasso (9.61) with regularization parameter $\lambda_n = 4\,B\,C\left\{\sqrt{\frac{\log d}{n}} + \delta\right\}$ for some $\delta \in (0, 1)$.

Corollary 9.26 *Consider a GLM satisfying conditions (G1) and (G2), the RSC condition* (9.62)*, and suppose the true regression vector θ^* is supported on a subset S of cardinality s. Given a sample size n large enough to ensure that $s\left\{\lambda_n^2 + \frac{\log d}{n}\right\} < \min\left\{\frac{4\kappa^2}{9}, \frac{\kappa}{64 c_1}\right\}$, any GLM Lasso solution $\widehat{\theta}$ satisfies the bounds*

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{9}{4}\frac{s\lambda_n^2}{\kappa^2} \quad \textit{and} \quad \|\widehat{\theta} - \theta^*\|_1 \leq \frac{6}{\kappa}\, s\, \lambda_n, \tag{9.63}$$

both with probability at least $1 - 2\,e^{-2n\delta^2}$.

We have already proved results of this form in Chapter 7 for the special case of the linear model; the proof here illustrates the application of our more general techniques.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

Proof Both results follow via an application of Corollary 9.20 with the subspaces

$$\mathbb{M}(S) = \bar{\mathbb{M}}(S) = \{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \notin S\}.$$

With this choice, note that we have $\Psi^2(\mathbb{M}) = s$; moreover, the assumed RSC condition (9.62) is a special case of our general definition with $\tau_n^2 = c_1 \frac{\log d}{n}$. In order to apply Corollary 9.20, we need to ensure that $\tau_n^2 \Psi^2(\mathbb{M}) < \frac{\kappa}{64}$, and since the local RSC holds over a ball with radius $R = 1$, we also need to ensure that $\frac{9}{4} \frac{\Psi^2(\mathbb{M}) \lambda_n^2}{\kappa^2} < 1$. Both of these conditions are guaranteed by our assumed lower bound on the sample size.

The only remaining step is to verify that the good event $\mathbb{G}(\lambda_n)$ holds with the probability stated in Corollary 9.26. Given the form (9.61) of the GLM log-likelihood, we can write the score function as the i.i.d. sum $\nabla \mathcal{L}_n(\theta^*) = \frac{1}{n} \sum_{i=1}^n V_i$, where $V_i \in \mathbb{R}^d$ is a zero-mean random vector with components

$$V_{ij} = \{\psi'(\langle x_i, \theta^* \rangle) - y_i\} \; x_{ij}.$$

Let us upper bound the moment generating function of these variables. For any $t \in \mathbb{R}$, we have

$$\begin{aligned}\log \mathbb{E}[e^{-tV_{ij}}] &= \log \mathbb{E}[e^{t y_i x_{ij}}] - t x_{ij} \psi'\big(\langle x_i, \theta^* \rangle\big) \\ &= \psi\big(t x_{ij} + \langle x_i, \theta^* \rangle\big) - \psi\big(\langle x_i, \theta^* \rangle\big) - t x_{ij} \psi'\big(\langle x_i, \theta^* \rangle\big).\end{aligned}$$

By a Taylor-series expansion, there is some intermediate $\tilde{t}$ such that

$$\log \mathbb{E}[e^{-tV_{ij}}] = \frac{1}{2} t^2 x_{ij}^2 \psi''\big(\tilde{t} x_{ij} + \langle x_i, \theta^* \rangle\big) \leq \frac{B^2 t^2 x_{ij}^2}{2},$$

where the final inequality follows from the boundedness condition (G2). Using independence of the samples, we have

$$\frac{1}{n} \log \mathbb{E}\left[e^{-t \sum_{i=1}^n V_{ij}}\right] \leq \frac{t^2 B^2}{2} \left(\frac{1}{n} \sum_{i=1}^n x_{ij}^2\right) \leq \frac{t^2 B^2 C^2}{2},$$

where the final step uses the column normalization (G1) on the columns of the design matrix $\mathbf{X}$. Since this bound holds for any $t \in \mathbb{R}$, we have shown that each element of the score function $\nabla \mathcal{L}_n(\theta^*) \in \mathbb{R}^d$ is zero-mean and sub-Gaussian with parameter at most $BC/\sqrt{n}$. Thus, sub-Gaussian tail bounds combined with the union bound guarantee that

$$\mathbb{P}\Big[\|\nabla \mathcal{L}_n(\theta^*)\|_\infty \geq t\Big] \leq 2 \exp\left(-\frac{n t^2}{2 B^2 C^2} + \log d\right).$$

Setting $t = 2BC \left\{\sqrt{\frac{\log d}{n}} + \delta\right\}$ completes the proof. □

9.5.3 *Bounds under ℓ_∞-curvature conditions*

The preceding results were devoted to error bounds in terms of quadratic-type norms, such as the Euclidean vector and Frobenius matrix norms. On the other hand, Theorem 9.24 provides bounds in terms of the dual norm Φ^*—that is, in terms of the ℓ_∞-norm in the case of ℓ_1-regularization. We now turn to exploration of such bounds in the case of generalized

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

linear models. As we discuss, ℓ_∞-bounds also lead to bounds in terms of the ℓ_2- and ℓ_1-norms, so that the resulting guarantees are in some sense stronger.

Recall that Theorem 9.24 is based on a restricted curvature condition (9.55). In the earlier Example 9.23, we discussed the specialization of this condition to the least-squares cost, and in Exercise 9.14, we work through the proof of an analogous result for generalized linear models with bounded cumulant generating functions ($\|\psi''\|_\infty \leq B$). More precisely, when the population cost satisfies an ℓ_∞-curvature condition over the ball $\mathbb{B}_2(R)$, and the covariates are i.i.d. and sub-Gaussian with parameter C, then the GLM log-likelihood $\mathcal{L}_n$ from equation (9.61) satisfies a bound of the form

$$\|\nabla\mathcal{L}_n(\theta^* + \Delta) - \nabla\mathcal{L}_n(\theta^*)\|_\infty \geq \kappa\|\Delta\|_\infty - \frac{c_0}{32}\sqrt{\frac{\log d}{n}}\|\Delta\|_1, \tag{9.64}$$

uniformly over $\mathbb{B}_\infty(1)$. Here is c_0 is a constant that depends only on the parameters (B, C).

Corollary 9.27 *In addition to the conditions of Corollary 9.26, suppose that the ℓ_∞-curvature condition* (9.64) *holds, and that the sample size is lower bounded as $n > c_0^2 s^2 \log d$. Then any optimal solution $\widehat{\theta}$ to the GLM Lasso* (9.61) *with regularization parameter $\lambda_n = 2\,BC\left(\sqrt{\frac{\log d}{n}} + \delta\right)$ satisfies*

$$\|\widehat{\theta} - \theta^*\|_\infty \leq 3\,\frac{\lambda_n}{\kappa} \tag{9.65}$$

with probability at least $1 - 2\,e^{-2n\delta^2}$.

Proof We prove this corollary by applying Theorem 9.24 with the familiar subspaces

$$\overline{\mathbb{M}}(S) = \mathbb{M}(S) = \{\theta \in \mathbb{R}^d \mid \theta_{S^c} = 0\},$$

for which we have $\Psi^2(\overline{\mathbb{M}}(S)) = s$. By assumption (9.64), the ℓ_∞-curvature condition holds with tolerance $\tau_n = \frac{c_0}{32}\sqrt{\frac{\log d}{n}}$, so that the condition $\tau_n\Psi^2(\mathbb{M}) < \frac{\kappa}{32}$ is equivalent to the lower bound $n > c_0^2\,s^2\,\log d$ on the sample size.

Since we have assumed the conditions of Corollary 9.26, we are guaranteed that the error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ satisfies the bound $\|\widehat{\Delta}\|_\infty \leq \|\widehat{\Delta}\|_2 \leq 1$ with high probability. This localization allows us to apply the local ℓ_∞-curvature condition to the error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$.

Finally, as shown in the proof of Corollary 9.26, if we choose the regularization parameter $\lambda_n = 2BC\left\{\sqrt{\frac{\log d}{n}} + \delta\right\}$, then the event $\mathbb{G}(\lambda_n)$ holds with probability at least $1 - e^{-2n\delta^2}$. We have thus verified that all the conditions needed to apply Theorem 9.24 are satisfied. □

The ℓ_∞-bound (9.65) is a stronger guarantee than our earlier bounds in terms of the ℓ_1- and ℓ_2-norms. For instance, under additional conditions on the smallest non-zero absolute values of θ^*, the ℓ_∞-bound (9.65) can be used to construct an estimator that has variable selection guarantees, which may not be possible with bounds in other norms. Moreover, as we explore

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

in Exercise 9.13, when combined with other properties of the error vector, Corollary 9.27 implies bounds on the ℓ_1- and ℓ_2-norm errors that are analogous to those in Corollary 9.26.

9.6 Bounds for group-structured sparsity

We now turn to the consequences of Theorem 9.19 for estimators based on the group Lasso penalty with non-overlapping groups, previously discussed in Example 9.3. For concreteness, we focus on the ℓ_2-version of the group Lasso penalty $\|\theta\|_{\mathcal{G},2} = \sum_{g \in \mathcal{G}} \|\theta_g\|_2$. As discussed in Example 9.6, one motivation for the group Lasso penalty are multivariate regression problems, in which the regression coefficients are assumed to appear on–off in a groupwise manner. The linear multivariate regression problem from Example 9.6 is the simplest example. In this section, we analyze the extension to generalized linear models. Accordingly, let us consider the *group GLM Lasso*

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \{\psi(\langle \theta,\, x_i \rangle) - y_i \langle \theta,\, x_i \rangle\} + \lambda_n \sum_{g \in \mathcal{G}} \|\theta_g\|_2 \right\}, \tag{9.66}$$

a family of estimators that includes the least-squares version of the group Lasso (9.14) as a particular case.

As with our previous corollaries, we assume that the samples $\{(x_i, y_i)\}_{i=1}^n$ are drawn i.i.d. from a generalized linear model (GLM) satisfying condition (G2). Letting $\mathbf{X}_g \in \mathbb{R}^{n \times |g|}$ denote the submatrix indexed by g, we also impose the following variant of condition (G1) on the design:

(G1′) The covariates satisfy the group normalization condition $\max_{g \in \mathcal{G}} \frac{|\!|\!|\mathbf{X}_g|\!|\!|_2}{\sqrt{n}} \leq C$.

Moreover, we assume an RSC condition of the form

$$\mathcal{E}_n(\Delta) \geq \kappa \|\Delta\|_2^2 - c_1 \left\{ \frac{m}{n} + \frac{\log |\mathcal{G}|}{n} \right\} \|\Delta\|_{\mathcal{G},2}^2 \qquad \text{for all } \|\Delta\|_2 \leq 1, \tag{9.67}$$

where m denotes the maximum size over all groups. As shown in Example 9.17 and Theorem 9.36, a lower bound of this form holds with high probability when the covariates $\{x_i\}_{i=1}^n$ are drawn i.i.d. from a zero-mean sub-Gaussian distribution. Our bound applies to any solution $\widehat{\theta}$ to the group GLM Lasso (9.66) based on a regularization parameter

$$\lambda_n = 4BC \left\{ \sqrt{\frac{m}{n}} + \sqrt{\frac{\log |\mathcal{G}|}{n}} + \delta \right\} \qquad \text{for some } \delta \in (0, 1).$$

Corollary 9.28 *Given n i.i.d. samples from a GLM satisfying conditions (G1′), (G2), the RSC condition* (9.67)*, suppose that the true regression vector* θ^* *has group support* $S_{\mathcal{G}}$*. As long as* $|S_{\mathcal{G}}| \left\{ \lambda_n^2 + \frac{m}{n} + \frac{\log |\mathcal{G}|}{n} \right\} < \min\left\{ \frac{4\kappa^2}{9}, \frac{\kappa}{64 c_1} \right\}$*, the estimate* $\widehat{\theta}$ *satisfies the bound*

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{9}{4} \frac{|S_{\mathcal{G}}| \lambda_n^2}{\kappa^2} \tag{9.68}$$

with probability at least $1 - 2e^{-2n\delta^2}$.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

In order to gain some intuition for this corollary, it is worthwhile to consider some special cases. The ordinary Lasso is a special case of the group Lasso, in which there are $|\mathcal{G}| = d$ groups, each of size $m = 1$. In this case, if we use the regularization parameter $\lambda_n = 8BC\sqrt{\frac{\log d}{n}}$, the bound (9.68) implies that

$$\|\widehat{\theta} - \theta^*\|_2 \precsim \frac{BC}{\kappa}\sqrt{\frac{|S_{\mathcal{G}}| \log d}{n}},$$

showing that Corollary 9.28 is a natural generalization of Corollary 9.26.

The problem of multivariate regression provides a more substantive example of the potential gains of using the group Lasso. Throughout this example, we take the regularization parameter $\lambda_n = 8BC\left\{\sqrt{\frac{m}{n}} + \sqrt{\frac{\log d}{n}}\right\}$ as given.

Example 9.29 (Faster rates for multivariate regression) As previously discussed in Example 9.6, the problem of multivariate regression is based on the linear observation model $\mathbf{Y} = \mathbf{Z}\mathbf{\Theta}^* + \mathbf{W}$, where $\mathbf{\Theta}^* \in \mathbb{R}^{p\times T}$ is a matrix of regression coefficients, $\mathbf{Y} \in \mathbb{R}^{n\times T}$ is a matrix of observations, and $\mathbf{W} \in \mathbb{R}^{n\times T}$ is a noise matrix. A natural group structure is defined by the rows of the regression matrix $\mathbf{\Theta}^*$, so that we have a total of p groups each of size T.

A naive approach would be to ignore the group sparsity, and simply apply the elementwise ℓ_1-norm as a regularizer to the matrix $\mathbf{\Theta}$. This set-up corresponds to a Lasso problem with $d = pT$ coefficients and elementwise sparsity $T|S_{\mathcal{G}}|$, so that Corollary 9.26 would guarantee an estimation error bound of the form

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\mathrm{F}} \precsim \sqrt{\frac{|S_{\mathcal{G}}|\, T \log(pT)}{n}}. \tag{9.69a}$$

By contrast, if we used the group Lasso estimator, which does explicitly model the grouping in the sparsity, then Corollary 9.28 would guarantee an error of the form

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\mathrm{F}} \precsim \sqrt{\frac{|S_{\mathcal{G}}|\, T}{n}} + \sqrt{\frac{|S_{\mathcal{G}}| \log p}{n}}. \tag{9.69b}$$

For $T > 1$, it can be seen that this error bound is always better than the Lasso error bound (9.69a), showing that the group Lasso is a better estimator when $\mathbf{\Theta}^*$ has a sparse group structure. In Chapter 15, we will develop techniques that can be used to show that the rate (9.69b) is the best possible for any estimator. Indeed, the two components in this rate have a very concrete interpretation: the first corresponds to the error associated with estimating $|S_{\mathcal{G}}|T$ parameters, assuming that the group structure is known. For $|S_{\mathcal{G}}| \ll p$, the second term is proportional to $\log\binom{p}{|S_{\mathcal{G}}|}$, and corresponds to the search complexity associated with finding the subset of $|S_{\mathcal{G}}|$ rows out of p that contain non-zero coefficients. ♣

We now turn to the proof of Corollary 9.28.

Proof We apply Corollary 9.20 using the model subspace $\mathbb{M}(S_{\mathcal{G}})$ defined in equation (9.24). From Definition 9.18 of the subspace constant with $\Phi(\theta) = \|\theta\|_{\mathcal{G},2}$, we have

$$\Psi(\mathbb{M}(S_{\mathcal{G}})) := \sup_{\theta\in\mathbb{M}(S_{\mathcal{G}})\setminus\{0\}} \frac{\sum_{g\in\mathcal{G}} \|\theta_g\|_2}{\|\theta\|_2} = \sqrt{|S_{\mathcal{G}}|}.$$

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

The assumed RSC condition (9.62) is a special case of our general definition with the tolerance parameter $\tau_n^2 = c_1 \left\{\frac{m}{n} + \frac{\log|\mathcal{G}|}{n}\right\}$ and radius $R = 1$. In order to apply Corollary 9.20, we need to ensure that $\tau_n^2 \Psi^2(\mathbb{M}) < \frac{\kappa}{64}$, and since the local RSC holds over a ball with radius $R = 1$, we also need to ensure that $\frac{9}{4} \frac{\Psi^2(\mathbb{M})\lambda_n^2}{\kappa^2} < 1$. Both of these conditions are guaranteed by our assumed lower bound on the sample size.

It remains to verify that, given the specified choice of regularization parameter λ_n, the event $\mathbb{G}(\lambda_n)$ holds with high probability.

Verifying the event $\mathbb{G}(\lambda_n)$: Using the form of the dual norm given in Table 9.1, we have $\Phi^*(\nabla\mathcal{L}_n(\theta^*)) = \max_{g\in\mathcal{G}} \|(\nabla\mathcal{L}_n(\theta^*))_g\|_2$. Based on the form of the GLM log-likelihood, we have $\nabla\mathcal{L}_n(\theta^*) = \frac{1}{n}\sum_{i=1}^n V_i$ where the random vector $V_i \in \mathbb{R}^d$ has components

$$V_{ij} = \Big\{\psi'(\langle x_i,\, \theta^*\rangle) - y_i\Big\}\, x_{ij}.$$

For each group g, we let $V_{i,g} \in \mathbb{R}^{|g|}$ denote the subvector indexed by elements of g. With this notation, we then have

$$\|(\nabla\mathcal{L}_n(\theta^*))_g\|_2 = \|\frac{1}{n}\sum_{i=1}^n V_{i,g}\|_2 = \sup_{u\in\mathbb{S}^{|g|-1}} \left\langle u,\, \frac{1}{n}\sum_{i=1}^n V_{i,g}\right\rangle,$$

where $\mathbb{S}^{|g|-1}$ is the Euclidean sphere in $\mathbb{R}^{|g|}$. From Example 5.8, we can find a $1/2$-covering of $\mathbb{S}^{|g|-1}$ in the Euclidean norm—say $\{u^1, \ldots, u^N\}$—with cardinality at most $N \leq 5^{|g|}$. By the standard discretization arguments from Chapter 5, we have

$$\|(\nabla\mathcal{L}_n(\theta^*))_g\|_2 \leq 2 \max_{j=1,\ldots,N} \left\langle u^j,\, \frac{1}{n}\sum_{i=1}^n V_{i,g}\right\rangle.$$

Using the same proof as Corollary 9.26, the random variable $\left\langle u^j,\, \frac{1}{n}\sum_{i=1}^n V_{i,g}\right\rangle$ is sub-Gaussian with parameter at most

$$\frac{B}{\sqrt{n}} \sqrt{\frac{1}{n}\sum_{i=1}^n \left\langle u^j,\, x_{i,g}\right\rangle^2} \leq \frac{BC}{\sqrt{n}},$$

where the inequality follows from condition (G1′). Consequently, from the union bound and standard sub-Gaussian tail bounds, we have

$$\mathbb{P}\Big[\|(\nabla\mathcal{L}_n(\theta^*))_g\|_2 \geq 2t\Big] \leq 2\exp\left(-\frac{nt^2}{2B^2C^2} + |g|\,\log 5\right).$$

Taking the union over all $|\mathcal{G}|$ groups yields

$$\mathbb{P}\Big[\max_{g\in\mathcal{G}} \|(\nabla\mathcal{L}_n(\theta^*))_g\|_2 \geq 2t\Big] \leq 2\exp\left(-\frac{nt^2}{2B^2C^2} + m\,\log 5 + \log|\mathcal{G}|\right),$$

where we have used the maximum group size m as an upper bound on each group size $|g|$. Setting $t^2 = \lambda_n^2$ yields the result. □

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

9.7 Bounds for overlapping decomposition-based norms

In this section, we turn to the analysis of the more "exotic" overlapping group Lasso norm, as previously introduced in Example 9.4. In order to motivate this estimator, let us return to the problem of multivariate regression.

Example 9.30 (Matrix decomposition in multivariate regression) Recall the problem of linear multivariate regression from Example 9.6: it is based on the linear observation model $\mathbf{Y} = \mathbf{Z}\mathbf{\Theta}^* + \mathbf{W}$, where $\mathbf{\Theta}^* \in \mathbb{R}^{p \times T}$ is an unknown matrix of regression coefficients. As discussed previously, the ordinary group Lasso is often applied in this setting, using the rows of the regression matrix to define the underlying set of groups. When the true regression matrix $\mathbf{\Theta}^*$ is actually row-sparse, then we can expect the group Lasso to yield a more accurate estimate than the usual elementwise Lasso: compare the bounds (9.69a) and (9.69b).

However, now suppose that we apply the group Lasso estimator to a problem for which the true regression matrix $\mathbf{\Theta}^*$ *violates* the row-sparsity assumption: concretely, let us suppose that $\mathbf{\Theta}^*$ has s total non-zero entries, each contained within a row of its own. In this setting, Corollary 9.28 guarantees a bound of the order

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\mathrm{F}} \precsim \sqrt{\frac{s\,T}{n}} + \sqrt{\frac{s \log p}{n}}. \tag{9.70}$$

However, if we were to apply the ordinary elementwise Lasso to this problem, then Corollary 9.26 would guarantee a bound of the form

$$|||\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*|||_{\mathrm{F}} \precsim \sqrt{\frac{s \log(pT)}{n}}. \tag{9.71}$$

This error bound is always smaller than the group Lasso bound (9.70), and substantially so for large T. Consequently, the ordinary group Lasso has the undesirable feature of being less statistically efficient than the ordinary Lasso in certain settings, despite its higher computational cost.

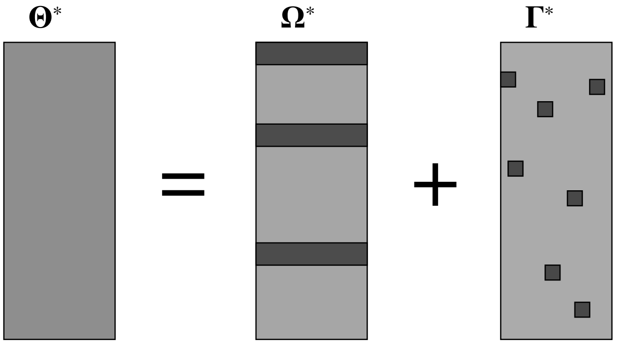


Figure 9.10 Illustration of the matrix decomposition norm (9.72) for the group Lasso applied to the matrix rows, combined with the elementwise ℓ_1-norm. The norm is defined by minimizing over all additive decompositions of $\mathbf{\Theta}^*$ as the sum of a row-sparse matrix $\mathbf{\Omega}^*$ with an elementwise-sparse matrix $\mathbf{\Gamma}^*$.

How do we remedy this issue? What would be desirable is an *adaptive estimator*, one that

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

achieves the ordinary Lasso rate (9.71) when the sparsity structure is elementwise, and the group Lasso rate (9.70) when the sparsity is row-wise. To this end, let us consider decomposing the regression matrix $\mathbf{\Theta}^*$ as a sum $\mathbf{\Omega}^* + \mathbf{\Gamma}^*$, where $\mathbf{\Omega}^*$ is a row-sparse matrix and $\mathbf{\Gamma}^*$ is elementwise sparse, as shown in Figure 9.10. Minimizing a weighted combination of the group Lasso and ℓ_1-norms over all such decompositions yields the norm

$$\Phi_\omega(\mathbf{\Theta}) = \inf_{\mathbf{\Omega}+\mathbf{\Gamma}=\mathbf{\Theta}} \left\{ \|\mathbf{\Gamma}\|_1 + \omega \sum_{j=1}^{p} \|\mathbf{\Omega}_j\|_2 \right\}, \tag{9.72}$$

which is a special case of the overlapping group Lasso (9.10). Our analysis to follow will show that an M-estimator based on such a regularizer exhibits the desired adaptivity. ♣

Let us return to the general setting, in which we view the parameter $\theta \in \mathbb{R}^d$ as a vector,[3] and consider the more general ℓ_1-plus-group overlap norm

$$\Phi_\omega(\theta) := \inf_{\alpha+\beta=\theta} \{\|\alpha\|_1 + \omega\|\beta\|_{\mathcal{G},2}\}, \tag{9.73}$$

where $\mathcal{G}$ is a set of disjoint groups, each of size at most m. The overlap norm (9.72) is a special case, where the groups are specified by the rows of the underlying matrix. For reasons to become clear in the proof, we use the weight

$$\omega := \frac{\sqrt{m} + \sqrt{\log|\mathcal{G}|}}{\sqrt{\log d}}. \tag{9.74}$$

With this set-up, the following result applies to the *adaptive group GLM Lasso*,

$$\widehat{\theta} \in \arg\min_{\theta\in\mathbb{R}^d} \Bigg\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n} \{\psi(\langle\theta,\, x_i\rangle) - \langle\theta,\, x_i y_i\rangle\}}_{\mathcal{L}_n(\theta)} + \lambda_n \Phi_\omega(\theta) \Bigg\}, \tag{9.75}$$

for which the Taylor-series error satisfies the RSC condition

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_2^2 - c_1 \frac{\log d}{n} \Phi_\omega^2(\Delta) \qquad \text{for all } \|\Delta\|_2 \leq 1. \tag{9.76}$$

Again, when the covariates $\{x_i\}_{i=1}^n$ are drawn i.i.d. from a zero-mean sub-Gaussian distribution, a bound of this form holds with high probability for any GLM (see Example 9.17 and Exercise 9.8).

With this set-up, the following result applies to any optimal solution $\widehat{\theta}$ of the adaptive group GLM Lasso (9.75) with $\lambda_n = 4BC(\sqrt{\frac{\log d}{n}} + \delta)$ for some $\delta \in (0, 1)$. Moreover, it supposes that the true regression vector can be decomposed as $\theta^* = \alpha^* + \beta^*$, where α is S_{elt}-sparse, and β^* is $S_{\mathcal{G}}$-group-sparse, and with $S_{\mathcal{G}}$ disjoint from S_{elt}.

[3] The problem of multivariate regression can be thought of as a particular case of the vector model with vector dimension $d = pT$, via the transformation $\mathbf{\Theta} \mapsto \mathrm{vec}(\mathbf{\Theta}) \in \mathbb{R}^{pT}$.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

Corollary 9.31 *Given n i.i.d. samples from a GLM satisfying conditions (G1′) and (G2), suppose that the RSC condition* (9.76) *with curvature $\kappa > 0$ holds, and that $\left\{\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_{\mathcal{G}}|}\right\}^2 \left\{\lambda_n^2 + \frac{\log d}{n}\right\} < \min\left\{\frac{\kappa^2}{36}, \frac{\kappa}{64c_1}\right\}$. Then the adaptive group GLM Lasso estimate $\widehat{\theta}$ satisfies the bounds*

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{36\lambda_n^2}{\kappa^2}\left\{\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_{\mathcal{G}}|}\right\}^2 \tag{9.77}$$

with probability at least $1 - 3e^{-8n\delta^2}$.

Remark: The most important feature of the bound (9.77) is its adaptivity to the elementwise versus group sparsity. This adaptivity stems from the fact that the choices of $S_{\mathcal{G}}$ and S_{elt} can be optimized so as to obtain the tightest possible bound, depending on the structure of the regression vector θ^*. To be concrete, consider the bound with the choice $\lambda_n = 8BC\sqrt{\frac{\log d}{n}}$. At one extreme, suppose that the true regression vector $\theta^* \in \mathbb{R}^d$ is purely elementwise sparse, in that each group contains at most one non-zero entry. In this case, we can apply the bound (9.77) with $S_{\mathcal{G}} = \emptyset$, leading to

$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim \frac{B^2\,C^2}{\kappa^2}\,\frac{s\,\log d}{n},$$

where $s = |S_{\text{elt}}|$ denotes the sparsity of θ^*. We thus recover our previous Lasso bound from Corollary 9.26 in this special case. At the other extreme, consider a vector that is “purely” group-sparse, in the sense that it has some subset of active groups $S_{\mathcal{G}}$, but no isolated sparse entries. The bound (9.77) with $S_{\text{elt}} = \emptyset$ then yields

$$\|\widehat{\theta} - \theta^*\|_2^2 \precsim \frac{B^2\,C^2}{\kappa^2}\left\{\frac{m|S_{\mathcal{G}}|}{n} + \frac{|S_{\mathcal{G}}|\log d}{n}\right\},$$

so that, in this special case, the decomposition method obtains the group Lasso rate from Corollary 9.28.

Let us now prove the corollary:

Proof In this case, we work through the details carefully, as the decomposability of the overlap norm needs some care. Recall the function $\mathcal{F}$ from equation (9.31), and let $\widehat{\Delta} = \widehat{\theta} - \theta^*$. Our proof is based on showing that any vector of the form $\Delta = t\widehat{\Delta}$ for some $t \in [0, 1]$ satisfies the bounds

$$\Phi_\omega(\Delta) \leq 4\left\{\sqrt{|S_{\text{elt}}|} + \omega\ \sqrt{|S_{\mathcal{G}}|}\right\}\|\Delta\|_2 \tag{9.78a}$$

and

$$\mathcal{F}(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_2^2 - c_1\,\frac{\log d}{n}\Phi_\omega^2(\Delta) - \frac{3\lambda_n}{2}\left\{\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_{\mathcal{G}}|}\right\}\ \|\Delta\|_2. \tag{9.78b}$$

Let us take these bounds as given for the moment, and then return to prove them. Substituting the bound (9.78a) into inequality (9.78b) and rearranging yields

$$\mathcal{F}(\Delta) \geq \|\Delta\|_2\left\{\kappa'\|\Delta\|_2 - \frac{3\lambda_n}{2}\left(\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_{\mathcal{G}}|}\right)\right\},$$

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

where $\kappa' := \frac{\kappa}{2} - 16c_1\frac{\log d}{n}\big(\sqrt{|S_{\text{elt}}|} + \omega\ \sqrt{|S_{\mathcal{G}}|}\big)^2$. Under the stated bound on the sample size n, we have $\kappa' \geq \frac{\kappa}{4}$, so that $\mathcal{F}$ is non-negative whenever

$$\|\Delta\|_2 \geq \frac{6\lambda_n}{\kappa}\Big(\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_{\mathcal{G}}|}\Big).$$

Finally, following through the remainder of the proof of Theorem 9.19 yields the claimed bound (9.77).

Let us now return to prove the bounds (9.78a) and (9.78b). To begin, a straightforward calculation shows that the dual norm is given by

$$\Phi^*_\omega(v) = \max\left\{\|v\|_\infty, \frac{1}{\omega}\max_{g\in\mathcal{G}}\|v_g\|_2\right\}.$$

Consequently, the event $\mathbb{G}(\lambda_n) := \{\Phi^*_\omega(\nabla\mathcal{L}_n(\theta^*)) \leq \frac{\lambda_n}{2}\}$ is equivalent to

$$\|\nabla\mathcal{L}_n(\theta^*)\|_\infty \leq \frac{\lambda_n}{2} \quad \text{and} \quad \max_{g\in\mathcal{G}}\|(\nabla\mathcal{L}_n(\theta^*))_g\|_2 \leq \frac{\lambda_n\omega}{2}. \tag{9.79}$$

We assume that these conditions hold for the moment, returning to verify them at the end of the proof.

Define $\Delta = t\widehat{\Delta}$ for some $t \in [0,1]$. Fix some decomposition $\theta^* = \alpha^* + \beta^*$, where α^* is S_{elt}-sparse and β^* is $S_{\mathcal{G}}$-group-sparse, and note that

$$\Phi_\omega(\theta^*) \leq \|\alpha^*\|_1 + \omega\|\beta^*\|_{\mathcal{G},2}.$$

Similarly, let us write $\Delta = \Delta_\alpha + \Delta_\beta$ for some pair such that

$$\Phi_\omega(\theta^* + \Delta) = \|\Delta_\alpha\|_1 + \omega\|\Delta_\beta\|_{\mathcal{G},2}.$$

Proof of inequality (9.78a)*:* Define the function

$$\mathcal{F}(\Delta) := \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) + \lambda_n\{\Phi_\omega(\theta^* + \Delta) - \Phi_\omega(\theta^*)\}.$$

Consider a vector of the form $\Delta = t\widehat{\Delta}$ for some scalar $t \in [0,1]$. Noting that $\mathcal{F}$ is convex and minimized at $\widehat{\Delta}$, we have

$$\mathcal{F}(\Delta) = \mathcal{F}(t\widehat{\Delta} + (1-t)0) \leq t\mathcal{F}(\widehat{\Delta}) + (1-t)\mathcal{F}(0) \leq \mathcal{F}(0).$$

Recalling that $\mathcal{E}_n(\Delta) = \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \langle\nabla\mathcal{L}_n(\theta^*), \Delta\rangle$, some algebra then leads to the inequality

$$\begin{aligned}
\mathcal{E}_n(\Delta) &\leq \Big|\langle\nabla\mathcal{L}_n(\theta^*), \Delta\rangle\Big| - \lambda_n\{\|\alpha^* + \Delta_\alpha\|_1 - \|\alpha^*\|_1\} - \lambda_n\omega\Big\{\|\beta^* + \Delta_\beta\|_{\mathcal{G},2} - \|\beta^*\|_{\mathcal{G},2}\Big\} \\
&\overset{\text{(i)}}{\leq} \Big|\langle\nabla\mathcal{L}_n(\theta^*), \Delta\rangle\Big| + \lambda_n\Big\{\|(\Delta_\alpha)_{S_{\text{elt}}}\|_1 - \|(\Delta_\alpha)_{S^c_{\text{elt}}}\|_1\Big\} + \lambda_n\omega\Big\{\|(\Delta_\beta)_{S_{\mathcal{G}}}\|_{\mathcal{G},2} - \|(\Delta_\beta)_{S_{\mathcal{G}}}\|_{\mathcal{G},2}\Big\} \\
&\overset{\text{(ii)}}{\leq} \frac{\lambda_n}{2}\Big\{3\|(\Delta_\alpha)_{S_{\text{elt}}}\|_1 - \|(\Delta_\alpha)_{S^c_{\text{elt}}}\|_1\Big\} + \frac{\lambda_n\omega}{2}\Big\{\|(\Delta_\beta)_{S_{\mathcal{G}}}\|_{\mathcal{G},2} - \|(\Delta_\beta)_{S_{\mathcal{G}}}\|_{\mathcal{G},2}\Big\}.
\end{aligned}$$

Here step (i) follows by decomposability of the ℓ_1 and the group norm, and step (ii) follows by using the inequalities (9.79). Since $\mathcal{E}_n(\Delta) \ge 0$ by convexity, rearranging yields

$$\begin{aligned}\|\Delta_\alpha\|_1 + \omega\|\Delta_\beta\|_{\mathcal{G},2} &\le 4\Big\{\|(\Delta_\alpha)_{S_{\text{elt}}}\|_1 + \omega\|(\Delta_\beta)_{S_{\mathcal{G}}}\|_{\mathcal{G},2}\Big\} \\ &\overset{\text{(iii)}}{\le} 4\Big\{\sqrt{|S_{\text{elt}}|}\;\|(\Delta_\alpha)_{S_{\text{elt}}}\|_2 + \omega\sqrt{|S_{\mathcal{G}}|}\;\|(\Delta_\beta)_{S_{\mathcal{G}}}\|_2\Big\} \\ &\le 4\Big\{\sqrt{|S_{\text{elt}}|} + \omega\sqrt{|S_{\mathcal{G}}|}\Big\}\Big\{\|(\Delta_\alpha)_{S_{\text{elt}}}\|_2 + \|(\Delta_\beta)_{S_{\mathcal{G}}}\|_2\Big\},\end{aligned} \tag{9.80}$$

where step (iii) follows from the subspace constants for the two decomposable norms. The overall vector Δ has the decomposition $\Delta = (\Delta_\alpha)_{S_{\text{elt}}} + (\Delta_\beta)_{S_{\mathcal{G}}} + \Delta_T$, where T is the complement of the indices in S_{elt} and $S_{\mathcal{G}}$. Noting that all three sets are disjoint by construction, we have

$$\|(\Delta_\alpha)_{S_{\text{elt}}}\|_2 + \|(\Delta_\beta)_{S_{\mathcal{G}}}\|_2 = \|(\Delta_\alpha)_{S_{\text{elt}}} + (\Delta_\beta)_{S_{\mathcal{G}}}\|_2 \le \|\Delta\|_2.$$

Combining with inequality (9.80) completes the proof of the bound (9.78a).

Proof of inequality (9.78b)*:* From the proof of Theorem 9.19, recall the lower bound (9.50). This inequality, combined with the RSC condition, guarantees that the function value $\mathcal{F}(\Delta)$ is at least

$$\begin{aligned}&\frac{\kappa}{2}\|\Delta\|_2^2 - c_1\frac{\log d}{n}\Phi_\omega^2(\Delta) - \big|\langle\nabla\mathcal{L}_n(\theta^*),\,\Delta\rangle\big| \\ &\quad + \lambda_n\{\|\alpha^* + \Delta_\alpha\|_1 - \|\alpha^*\|_1\} + \lambda_n\omega\{\|\beta^* + \Delta_\beta\|_{\mathcal{G},2} - \|\beta^*\|_{\mathcal{G},2}\}.\end{aligned}$$

Again, applying the dual norm bounds (9.79) and exploiting decomposability leads to the lower bound (9.78b).

Verifying inequalities (9.79)*:* The only remaining detail is to verify that the conditions (9.79) defining the event $\mathbb{G}(\lambda_n)$. From the proof of Corollary 9.26, we have

$$\mathbb{P}[\|\nabla\mathcal{L}_n(\theta^*)\|_\infty \ge t] \le d\,e^{-\frac{nt^2}{2B^2C^2}}.$$

Similarly, from the proof of Corollary 9.28, we have

$$\mathbb{P}\Big[\frac{1}{\omega}\max_{g\in\mathcal{G}}\|(\nabla\mathcal{L}_n(\theta^*))_g\|_2 \ge 2t\Big] \le 2\exp\left(-\frac{n\omega^2t^2}{2B^2C^2} + m\,\log 5 + \log|\mathcal{G}|\right).$$

Setting $t = 4BC\left\{\sqrt{\frac{\log d}{n}} + \delta\right\}$ and performing some algebra yields the claimed lower bound $\mathbb{P}[\mathbb{G}(\lambda_n)] \ge 1 - 3e^{-8n\delta^2}$. □

9.8 Techniques for proving restricted strong convexity

All of the previous results rely on the empirical cost function satisfying some form of restricted curvature condition. In this section, we turn to a deeper investigation of the conditions under which restricted strong convexity conditions, as previously formalized in Definition 9.15, are satisfied.

Before proceeding, let us set up some notation. Given a collection of samples $Z_1^n = \{Z_i\}_{i=1}^n$, we write the empirical cost as $\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(\theta; Z_i)$, where $\mathcal{L}$ is the loss applied to a single sample. We can then define the error in the first-order Taylor expansion of $\mathcal{L}$ for sample Z_i, namely

$$\mathcal{E}(\Delta\,; Z_i) := \mathcal{L}(\theta^* + \Delta; Z_i) - \mathcal{L}(\theta^*; Z_i) - \langle \nabla \mathcal{L}(\theta^*; Z_i),\, \Delta \rangle\,.$$

By construction, we have $\mathcal{E}_n(\Delta) = \frac{1}{n}\sum_{i=1}^n \mathcal{E}(\Delta\,; Z_i)$. Given the population cost function $\bar{\mathcal{L}}(\theta) := \mathbb{E}[\mathcal{L}_n(\theta; Z_1^n)]$, a local form of strong convexity can be defined in terms of its Taylor-series error

$$\bar{\mathcal{E}}(\Delta) := \bar{\mathcal{L}}(\theta^* + \Delta) - \bar{\mathcal{L}}(\theta^*) - \Big\langle \nabla \bar{\mathcal{L}}(\theta^*),\, \Delta \Big\rangle\,. \tag{9.81}$$

We say that the population cost is (locally) *κ-strongly convex* around the minimizer θ^* if there exists a radius $R > 0$ such that

$$\bar{\mathcal{E}}(\Delta) \geq \kappa \|\Delta\|_2^2 \qquad \text{for all } \Delta \in \mathbb{B}_2(R) := \{\Delta \in \Omega \mid \|\theta\|_2 \leq R\}. \tag{9.82}$$

We wish to see when this type of curvature condition is inherited by the sample-based error $\mathcal{E}_n(\Delta)$. At a high level, then, our goal is clear: in order to establish a form of restricted strong convexity (RSC), we need to derive some type of uniform law of large numbers (see Chapter 4) for the zero-mean stochastic process

$$\Big\{\mathcal{E}_n(\Delta) - \bar{\mathcal{E}}(\Delta),\ \Delta \in \mathbb{S}\Big\}, \tag{9.83}$$

where $\mathbb{S}$ is a suitably chosen subset of $\mathbb{B}_2(R)$.

Example 9.32 (Least squares) To gain intuition in a specific example, recall the quadratic cost function $\mathcal{L}(\theta; y_i, x_i) = \frac{1}{2}(y - \langle \theta,\, x_i\rangle)^2$ that underlies least-squares regression. In this case, we have $\mathcal{E}(\Delta\,; x_i, y_i) = \frac{1}{2}\,\langle \Delta,\, x_i\rangle^2$, and hence

$$\mathcal{E}_n(\Delta) = \frac{1}{2n}\sum_{i=1}^n \langle \Delta,\, x_i\rangle^2 = \frac{1}{2n}\|\mathbf{X}\Delta\|_2^2,$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the usual design matrix. Denoting $\mathbf{\Sigma} = \mathrm{cov}(x)$, we find that

$$\bar{\mathcal{E}}(\Delta) = \mathbb{E}[\mathcal{E}_n(\Delta)] = \tfrac{1}{2}\Delta^{\mathrm{T}}\mathbf{\Sigma}\Delta.$$

Thus, our specific goal in this case is to establish a uniform law for the family of random variables

$$\Big\{\frac{1}{2}\Delta^{\mathrm{T}}\Big(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n} - \mathbf{\Sigma}\Big)\Delta,\ \Delta \in \mathbb{S}\Big\}. \tag{9.84}$$

When $\mathbb{S} = \mathbb{B}_2(1)$, the supremum over this family is equal to the operator norm $|||\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n} - \mathbf{\Sigma}|||_2$, a quantity that we studied in Chapter 6. When $\mathbb{S}$ involves an additional ℓ_1-constraint, then a uniform law over this family amounts to establishing a restricted eigenvalue condition, as studied in Chapter 7. ♣

9.8.1 *Lipschitz cost functions and Rademacher complexity*

This section is devoted to showing how the problem of establishing RSC for Lipschitz cost functions can be reduced to controlling a version of the Rademacher complexity. As the reader might expect, the symmetrization and contraction techniques from Chapter 4 turn out to be useful.

We say that $\mathcal{L}$ is locally *L-Lipschitz* over the ball $\mathbb{B}_2(R)$ if for each sample $Z = (x, y)$

$$\left|\mathcal{L}(\theta; Z) - \mathcal{L}(\widetilde{\theta}; Z)\right| \leq L \left|\langle \theta,\ x\rangle - \left\langle \widetilde{\theta},\ x\right\rangle\right| \qquad \text{for all } \theta, \widetilde{\theta} \in \mathbb{B}_2(R). \tag{9.85}$$

Let us illustrate this definition with an example.

Example 9.33 (Cost functions for binary classification) The class of Lipschitz cost functions includes various objective functions for binary classification, in which the goal is to use the covariates $x \in \mathbb{R}^d$ to predict an underlying class label $y \in \{-1, 1\}$. The simplest approach is based on a linear classification rule: given a weight vector $\theta \in \mathbb{R}^d$, the sign of the inner product $\langle\theta,\ x\rangle$ is used to make decisions. If we disregard computational issues, the most natural cost function is the 0–1 cost $\mathbb{I}[y\,\langle\theta,\ x\rangle < 0]$, which assigns a penalty of 1 if the decision is incorrect, and returns 0 otherwise. (Note that $y\,\langle\theta,\ x\rangle < 0$ if and only if $\text{sign}(\langle\theta,\ x\rangle) \neq y$.)

For instance, the *logistic cost* takes the form

$$\mathcal{L}(\theta; (x, y)) := \log(1 + e^{\langle\theta,\, x\rangle}) - y\,\langle\theta,\ x\rangle, \tag{9.86}$$

and it is straightforward to verify that this cost function satisfies the Lipschitz condition with $L = 2$. Similarly, the support vector machine approach to classification is based on the *hinge cost*

$$\mathcal{L}(\theta; (x, y)) := \max\left\{0, 1 - y\,\langle\theta,\ x\rangle\right\} \equiv \left(1 - y\,\langle\theta,\ x\rangle\right)_+, \tag{9.87}$$

which is Lipschitz with parameter $L = 1$. Note that the least-squares cost function $\mathcal{L}(\theta; (x, y)) = \frac{1}{2}(y - \langle\theta,\ x\rangle)^2$ is *not* Lipschitz unless additional boundedness conditions are imposed. A similar observation applies to the exponential cost function $\mathcal{L}(\theta; (x, y)) = e^{-y\langle\theta,\, x\rangle}$. ♣

In this section, we prove that Lipschitz functions with regression-type data $z = (x, y)$ satisfy a certain form of restricted strong convexity, depending on the tail fluctuations of the covariates. The result itself involves a complexity measure associated with the norm ball of the regularizer Φ. More precisely, letting $\{\varepsilon_i\}_{i=1}^n$ be an i.i.d. sequence of Rademacher variables, we define the symmetrized random vector $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i$, and the random variable

$$\Phi^*(\bar{x}_n) := \sup_{\Phi(\theta)\leq 1} \left\langle \theta,\ \frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i \right\rangle. \tag{9.88}$$

When $x_i \sim \mathcal{N}(0, \mathbf{I}_d)$, the mean $\mathbb{E}[\Phi^*(\bar{x}_n)]$ is proportional to the Gaussian complexity of the unit ball $\{\theta \in \mathbb{R}^d \mid \Phi(\theta) \leq 1\}$. (See Chapter 5 for an in-depth discussion of the Gaussian complexity and its properties.) More generally, the quantity (9.88) reflects the size of the Φ-unit ball with respect to the fluctuations of the covariates.

The following theorem applies to any norm Φ that dominates the Euclidean norm, in the sense that $\Phi(\Delta) \geq \|\Delta\|_2$ uniformly. For a pair of radii $0 < R_\ell < R_u$, it guarantees a form of restricted strong convexity over the “donut” set

$$\mathbb{B}_2(R_\ell, R_u) := \{\Delta \in \mathbb{R}^d \mid R_\ell \leq \|\Delta\|_2 \leq R_u\}. \tag{9.89}$$

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

The high-probability statement is stated in terms of the random variable $\Phi^*(\bar{x}_n)$, as well as the quantity $M_n(\Phi; R) := 4\log\left(\frac{R_u}{R_\ell}\right)\ \log \sup_{\theta\neq 0}\left(\frac{\Phi(\theta)}{\|\theta\|_2}\right)$, which arises for technical reasons.

Theorem 9.34 *Suppose that the cost function $\mathcal{L}$ is L-Lipschitz* (9.85)*, and the population cost $\bar{\mathcal{L}}$ is locally κ-strongly convex* (9.82) *over the ball $\mathbb{B}_2(R_u)$. Then for any $\delta > 0$, the first-order Taylor error $\mathcal{E}_n$ satisfies*

$$\left|\mathcal{E}_n(\Delta) - \bar{\mathcal{E}}(\Delta)\right| \leq 16L\,\Phi(\Delta)\,\delta \qquad \textit{for all } \Delta \in \mathbb{B}_2(R_\ell, R_u) \tag{9.90}$$

with probability at least $1 - M_n(\Phi; R)\ \inf_{\lambda>0} \mathbb{E}[e^{\lambda(\Phi^*(\bar{x}_n)-\delta)}]$.

For Lipschitz functions, this theorem reduces the question of establishing RSC to that of controlling the random variable $\Phi^*(\bar{x}_n)$. Let us consider a few examples to illustrate the consequences of Theorem 9.34.

Example 9.35 (Lipschitz costs and group Lasso) Consider the group Lasso norm $\Phi(\theta) = \sum_{g\in\mathcal{G}} \|\theta_g\|_2$, where we take groups of equal size m for simplicity. Suppose that the covariates $\{x_i\}_{i=1}^n$ are drawn i.i.d. as $\mathcal{N}(0, \mathbf{\Sigma})$ vectors, and let $\sigma^2 = |||\mathbf{\Sigma}|||_2$. In this case, we show that for any L-Lipschitz cost function, the inequality

$$\left|\mathcal{E}_n(\Delta) - \bar{\mathcal{E}}(\Delta)\right| \leq 16L\sigma\left\{\sqrt{\frac{m}{n}} + \sqrt{\frac{2\log|\mathcal{G}|}{n}} + \delta\right\} \sum_{g\in\mathcal{G}} \|\Delta_g\|_2$$

holds uniformly for all $\Delta \in \mathbb{B}_2(\frac{1}{d}, 1)$ with probability at least $1 - 4\log^2(d)\, e^{-\frac{n\delta^2}{2}}$.

In order to establish this claim, we begin by noting that $\Phi^*(\bar{x}_n) = \max_{g\in\mathcal{G}} \|(\bar{x}_n)_g\|_2$ from Table 9.1. Consequently, we have

$$\mathbb{E}[e^{\lambda\Phi^*(\bar{x}_n)}] \leq \sum_{g\in\mathcal{G}} \mathbb{E}\left[e^{\lambda(\|(\bar{x}_n)_g\|_2)}\right] = \sum_{g\in\mathcal{G}} \mathbb{E}\left[e^{\lambda(\|(\bar{x}_n)_g\|_2 - \mathbb{E}[\|(\bar{x}_n)_g\|_2])}\right]\ e^{\lambda\mathbb{E}[\|(\bar{x}_n)_g\|_2]}.$$

By Theorem 2.26, the random variable $\|(\bar{x}_n)_g\|_2$ has sub-Gaussian concentration around its mean with parameter $\sigma/\sqrt{n}$, whence $\mathbb{E}[e^{\lambda(\|(\bar{x}_n)_g\|_2 - \mathbb{E}[\|(\bar{x}_n)_g\|_2])}] \leq e^{\frac{\lambda^2\sigma^2}{2n}}$. By Jensen's inequality, we have

$$\mathbb{E}[\|(\bar{x}_n)_g\|_2] \leq \sqrt{\mathbb{E}[\|(\bar{x}_n)_g\|_2^2]} \leq \sigma\sqrt{\frac{m}{n}},$$

using the fact that $\sigma^2 = |||\mathbf{\Sigma}|||_2$. Putting together the pieces, we have shown that

$$\inf_{\lambda>0} \log \mathbb{E}[e^{\lambda(\Phi^*(\bar{x}_n)-(\epsilon+\sigma\sqrt{\frac{m}{n}}))}] \leq \log|\mathcal{G}| + \inf_{\lambda>0}\left\{\frac{\lambda^2\sigma^2}{2n} - \lambda\epsilon\right\} = \log|\mathcal{G}| - \frac{n\epsilon^2}{2\sigma^2}.$$

With the choices $R_u = 1$ and $R_\ell = \frac{1}{d}$, we have

$$M_n(\Phi; R) = 4\log(d)\ \log|\mathcal{G}| \leq 4\log^2(d),$$

since $|\mathcal{G}| \leq d$. Thus, setting $\epsilon = 2\sigma\left\{\sqrt{\frac{\log|\mathcal{G}|}{n}} + \epsilon\right\}$ and applying Theorem 9.34 yields the stated claim. ♣

In Chapter 10, we discuss some consequences of Theorem 9.34 for estimating low-rank matrices. Let us now turn to its proof.

Proof Recall that

$$\mathcal{E}(\Delta; z_i) := \mathcal{L}(\theta^* + \Delta; z_i) - \mathcal{L}(\theta^*; z_i) - \langle \nabla \mathcal{L}(\theta^*; z_i),\, \Delta \rangle$$

denotes the Taylor-series error associated with a single sample $z_i = (x_i, y_i)$.

Showing the Taylor error is Lipschitz: We first show that $\mathcal{E}$ is a $2L$-Lipschitz function in $\langle \Delta,\, x_i \rangle$. To establish this claim, note if that we let $\frac{\partial \mathcal{L}}{\partial u}$ denote the derivative of $\mathcal{L}$ with respect to $u = \langle \theta,\, x \rangle$, then the Lipschitz condition implies that $\|\frac{\partial \mathcal{L}}{\partial u}\|_\infty \leq L$. Consequently, by the chain rule, for any sample $z_i \in \mathcal{Z}$ and parameters $\Delta, \widetilde{\Delta} \in \mathbb{R}^d$, we have

$$\left|\left\langle \nabla \mathcal{L}(\theta^*; Z_i),\, \Delta - \widetilde{\Delta} \right\rangle\right| \leq \left|\frac{\partial \mathcal{L}}{\partial u}(\theta^*; Z_i)\right| \left|\langle \Delta,\, x_i \rangle - \left\langle \widetilde{\Delta},\, x_i \right\rangle\right| \leq L \left|\langle \Delta,\, x_i \rangle - \left\langle \widetilde{\Delta},\, x_i \right\rangle\right|. \tag{9.91}$$

Putting together the pieces, for any pair $\Delta, \widetilde{\Delta}$, we have

$$\begin{aligned} \left|\mathcal{E}(\Delta; Z_i) - \mathcal{E}(\widetilde{\Delta}; Z_i)\right| &\leq \left|\mathcal{L}(\theta^* + \Delta; Z_i) - \mathcal{L}(\theta^* + \widetilde{\Delta}; Z_i)\right| + \left|\left\langle \nabla \mathcal{L}(\theta^*; Z_i),\, \Delta - \widetilde{\Delta} \right\rangle\right| \\ &\leq 2L \left|\langle \Delta,\, x_i \rangle - \left\langle \widetilde{\Delta},\, x_i \right\rangle\right|, \end{aligned} \tag{9.92}$$

where the second inequality applies our Lipschitz assumption, and the gradient bound (9.91). Thus, the Taylor error is a $2L$-Lipschitz function in $\langle \Delta,\, x_i \rangle$.

Tail bound for fixed radii: Next we control the difference $|\mathcal{E}_n(\Delta) - \bar{\mathcal{E}}(\Delta)|$ uniformly over certain sets defined by fixed radii. More precisely, for positive quantities (r_1, r_2), define the set

$$\mathbb{C}(r_1, r_2) := \mathbb{B}_2(r_2) \cap \{\Phi(\Delta) \leq r_1 \, \|\Delta\|_2\},$$

and the random variable $A_n(r_1, r_2) := \frac{1}{4 r_1 r_2 L} \sup_{\Delta \in \mathbb{C}(r_1, r_2)} \left|\mathcal{E}_n(\Delta) - \bar{\mathcal{E}}(\Delta)\right|$. The choice of radii can be implicitly understood, so that we adopt the shorthand A_n.

Our goal is to control the probability of the event $\{A_n \geq \delta\}$, and we do so by controlling the moment generating function. By our assumptions, the Taylor error has the additive decomposition $\mathcal{E}_n(\Delta) = \frac{1}{n} \sum_{i=}^n \mathcal{E}(\Delta; Z_i)$. Thus, letting $\{\varepsilon_i\}_{i=1}^n$ denote an i.i.d. Rademacher sequence, applying the symmetrization upper bound from Proposition 4.11(b) yields

$$\mathbb{E}[e^{\lambda A_n}] \leq \mathbb{E}_{Z,\varepsilon}\left[\exp\left(2\lambda \sup_{\Delta \in \mathbb{C}(r_1, r_2)} \left|\frac{1}{4\, L r_1\, r_2}\, \frac{1}{n} \sum_{i=1}^n \varepsilon_i \, \mathcal{E}(\Delta; Z_i)\right|\right)\right].$$

Now we have

$$\mathbb{E}[e^{\lambda A_n}] \overset{\text{(i)}}{\leq} \mathbb{E}\left[\exp\left(\frac{\lambda}{r_1 r_2} \sup_{\Delta \in \mathbb{C}(r_1, r_2)} \left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i \, \langle \Delta,\, x_i \rangle\right|\right)\right] \overset{\text{(ii)}}{\leq} \mathbb{E}\left[\exp\left\{\lambda \, \Phi^*\left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i x_i\right)\right\}\right],$$

where step (i) uses the Lipschitz property (9.92) and the Ledoux–Talagrand contraction inequality (5.61), whereas step (ii) follows from applying Hölder's inequality to the regularizer and its dual (see Exercise 9.7), and uses the fact that $\Phi^*(\Delta) \leq r_1 r_2$ for any vector

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

$\Delta \in \mathbb{C}(r_1, r_2)$. Adding and subtracting the scalar $\delta > 0$ then yields

$$\log \mathbb{E}[e^{\lambda(A_n - \delta)}] \le -\lambda\delta + \log \mathbb{E}\left[\exp\left\{\lambda\, \Phi^*\Big(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i x_i\Big)\right\}\right],$$

and consequently, by Markov's inequality,

$$\mathbb{P}[A_n(r_1, r_2) \ge \delta] \le \inf_{\lambda>0} \mathbb{E}\left[\exp\Big(\lambda\{\Phi^*(\bar{x}_n) - \delta\}\Big)\right]. \tag{9.93}$$

Extension to uniform radii via peeling: This bound (9.93) applies to fixed choice of quantities (r_1, r_2), whereas the claim of Theorem 9.34 applies to possibly random choices—namely, $\frac{\Phi(\Delta)}{\|\Delta\|_2}$ and $\|\Delta\|_2$, respectively, where Δ might be chosen in a way dependent on the data. In order to extend the bound to all choices, we make use of a peeling argument.

Let $\mathcal{E}$ be the event that the bound (9.90) is violated. For positive integers (k, ℓ), define the sets

$$\mathbb{S}_{k,\ell} := \left\{\Delta \in \mathbb{R}^d \mid 2^{k-1} \le \frac{\Phi(\Delta)}{\|\Delta\|_2} \le 2^k \text{ and } 2^{\ell-1} R_\ell \le \|\Delta\|_2 \le 2^\ell R_\ell\right\}.$$

By construction, any vector that can possibly violate the bound (9.90) is contained in the union $\bigcup_{k=1}^{N_1} \bigcup_{\ell=1}^{N_2} \mathbb{S}_{k,\ell}$, where $N_1 := \lceil \log \sup_{\theta \neq 0} \frac{\Phi(\theta)}{\|\theta\|} \rceil$ and $N_2 := \lceil \log \frac{R_u}{R_\ell} \rceil$. Suppose that the bound (9.90) is violated by some $\widehat{\Delta} \in \mathbb{S}_{k,\ell}$. In this case, we have

$$\left|\mathcal{E}_n(\widehat{\Delta}) - \bar{\mathcal{E}}(\widehat{\Delta})\right| \ge 16L \frac{\Phi(\widehat{\Delta})}{\|\widehat{\Delta}\|_2} \|\widehat{\Delta}\|_2\, \delta \ge 16L 2^{k-1} 2^{\ell-1} R_\ell\, \delta = 4L 2^k 2^\ell R_\ell\, \delta,$$

which implies that $A_n(2^k, 2^\ell R_\ell) \ge \delta$. Consequently, we have shown that

$$\mathbb{P}[\mathcal{E}] \le \sum_{k=1}^{N_1} \sum_{\ell=1}^{N_2} \mathbb{P}[A_n(2^k, 2^\ell R_\ell) \ge \delta] \le N_1 N_2 \inf_{\lambda>0} \mathbb{E}[e^{\lambda(\Phi^*(\bar{x}_n) - \delta)}],$$

where the final step follows by the union bound, and the tail bound (9.93). Given the upper bound $N_1 N_2 \le 4 \log(\sup_{\theta\neq 0} \frac{\Phi(\theta)}{\|\theta\|}) \log(\frac{R_u}{R_\ell}) = M_n(\Phi; R)$, the claim follows. □

9.8.2 A one-sided bound via truncation

In the previous section, we actually derived two-sided bounds on the difference between the empirical $\mathcal{E}_n$ and population $\bar{\mathcal{E}}$ form of the Taylor-series error. The resulting upper bounds on $\mathcal{E}_n$ guarantee a form of *restricted smoothness*, one which is useful in proving fast convergence rates of optimization algorithms. (See the bibliographic section for further details.) However, for proving bounds on the estimation error, as has been our focus in this chapter, it is only *restricted strong convexity*—that is, the lower bound on the Taylor-series error—that is required.

In this section, we show how a truncation argument can be used to derive restricted strong convexity for generalized linear models. Letting $\{\varepsilon_i\}_{i=1}^n$ denote an i.i.d. sequence of Rademacher variables, we define a complexity measure involving the dual norm Φ^*—namely

$$\mu_n(\Phi^*) := \mathbb{E}_{x,\varepsilon}\left[\Phi^*\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i\right)\right] = \mathbb{E}\left[\sup_{\Phi(\Delta)\leq 1} \frac{1}{n}\sum_{i=1}^n \varepsilon_i \langle \Delta,\, x_i\rangle\right].$$

This is simply the Rademacher complexity of the linear function class $x \mapsto \langle \Delta,\, x\rangle$ as Δ ranges over the unit ball of the norm Φ.

Our theory applies to covariates $\{x_i\}_{i=1}^n$ drawn i.i.d. from a zero-mean distribution such that, for some positive constants (α, β), we have

$$\mathbb{E}\Big[\langle \Delta,\, x\rangle^2\Big] \geq \alpha \quad \text{and} \quad \mathbb{E}\Big[\langle \Delta,\, x\rangle^4\Big] \leq \beta \qquad \text{for all vectors } \Delta \in \mathbb{R}^d \text{ with } \|\Delta\|_2 = 1. \tag{9.94}$$

Theorem 9.36 *Consider any generalized linear model with covariates drawn from a zero-mean distribution satisfying the condition* (9.94). *Then the Taylor-series error $\mathcal{E}_n$ in the log-likelihood is lower bounded as*

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_2^2 - c_0\, \mu_n^2(\Phi^*)\, \Phi^2(\Delta) \qquad \textit{for all } \Delta \in \mathbb{R}^d \textit{ with } \|\Delta\|_2 \leq 1 \tag{9.95}$$

with probability at least $1 - c_1 e^{-c_2 n}$.

In this statement, the constants (κ, c_0, c_1, c_2) can depend on the GLM, the fixed vector θ^* and (α, β), but are independent of dimension, sample size, and regularizer.

Proof Using a standard formula for the remainder in the Taylor series, we have

$$\mathcal{E}_n(\Delta) = \frac{1}{n}\sum_{i=1}^n \psi''\Big(\langle \theta^*,\, x_i\rangle + t\,\langle \Delta,\, x_i\rangle\Big)\langle \Delta,\, x_i\rangle^2,$$

for some scalar $t \in [0, 1]$. We proceed via a truncation argument. Fix some vector $\Delta \in \mathbb{R}^d$ with Euclidean norm $\|\Delta\|_2 = \delta \in (0, 1]$, and set $\tau = K\delta$ for a constant $K > 0$ to be chosen. Since the function $\varphi_\tau(u) = u^2\mathbb{I}[|u| \leq 2\tau]$ lower bounds the quadratic and ψ'' is positive, we have

$$\mathcal{E}_n(\Delta) \geq \frac{1}{n}\sum_{i=1}^n \psi''\Big(\langle \theta^*,\, x_i\rangle + t\,\langle \Delta,\, x_i\rangle\Big)\, \varphi_\tau(\langle \Delta,\, x_i\rangle)\, \mathbb{I}[|\langle \theta^*,\, x_i\rangle| \leq T], \tag{9.96}$$

where T is a second truncation parameter to be chosen. Since φ_τ vanishes outside the interval $[-2\tau, 2\tau]$ and $\tau \leq K$, for any positive term in this sum, the absolute value $|\langle \theta^*,\, x_i\rangle + t\,\langle \Delta,\, x_i\rangle|$ is at most $T + 2K$, and hence

$$\mathcal{E}_n(\Delta) \geq \gamma\, \frac{1}{n}\sum_{i=1}^n \varphi_\tau\,(\langle \Delta,\, x_i\rangle)\;\, \mathbb{I}[|\langle \theta^*,\, x_i\rangle| \leq T] \qquad \text{where } \gamma := \min_{|u|\leq T+2K} \psi''(u).$$

Based on this lower bound, it suffices to show that for all $\delta \in (0, 1]$ and for $\Delta \in \mathbb{R}^d$ with $\|\Delta\|_2 = \delta$, we have

$$\frac{1}{n}\sum_{i=1}^n \varphi_{\tau(\delta)}(\langle \Delta,\, x_i\rangle)\, \mathbb{I}[|\langle \theta^*,\, x_i\rangle| \leq T] \geq c_3\delta^2 - c_4\mu_n(\Phi^*)\Phi(\Delta)\,\delta. \tag{9.97}$$

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

When this bound holds, then inequality (9.95) holds with constants (κ, c_0) depending on (c_3, c_4, γ). Moreover, we claim that the problem can be reducing to proving the bound (9.97) for $\delta = 1$. Indeed, given any vector with Euclidean norm $\|\Delta\|_2 = \delta > 0$, we can apply the bound (9.97) to the rescaled unit-norm vector Δ/δ to obtain

$$\frac{1}{n}\sum_{i=1}^{n} \varphi_{\tau(1)}(\langle \Delta/\delta,\, x_i\rangle)\ \mathbb{I}[|\langle \theta^*,\, x_i\rangle| \leq T] \geq c_3\left\{1 - c_4\mu_n(\Phi^*)\frac{\Phi(\Delta)}{\delta}\right\},$$

where $\tau(1) = K$, and $\tau(\delta) = K\delta$. Noting that $\varphi_{\tau(1)}(u/\delta) = (1/\delta)^2\varphi_{\tau(\delta)}(u)$, the claim follows by multiplying both sides by δ^2. Thus, the remainder of our proof is devoted to proving (9.97) with $\delta = 1$. In fact, in order to make use of a contraction argument for Lipschitz functions, it is convenient to define a new truncation function

$$\widetilde{\varphi}_\tau(u) = u^2\ \mathbb{I}[|u| \leq \tau] + (u - 2\tau)^2\ \mathbb{I}[\tau < u \leq 2\tau] + (u + 2\tau)^2\ \mathbb{I}[-2\tau \leq u < -\tau].$$

Note that it is Lipschitz with parameter 2τ. Since $\widetilde{\varphi}_\tau$ lower bounds φ_τ, it suffices to show that for all unit-norm vectors Δ, we have

$$\frac{1}{n}\sum_{i=1}^{n} \widetilde{\varphi}_\tau\Big(\langle \Delta,\, x_i\rangle\Big)\ \mathbb{I}[|\langle \theta^*,\, x_i\rangle| \leq T] \geq c_3 - c_4\mu_n(\Phi^*)\Phi(\Delta). \tag{9.98}$$

For a given radius $r \geq 1$, define the random variable

$$Z_n(r) := \sup_{\substack{\|\Delta\|_2=1\\ \Phi(\Delta)\leq r}} \left|\frac{1}{n}\sum_{i=1}^{n} \widetilde{\varphi}_\tau(\langle \Delta,\, x_i\rangle)\ \mathbb{I}[|\langle \theta^*,\, x_i\rangle| \leq T] - \mathbb{E}[\widetilde{\varphi}_\tau(\langle \Delta,\, x\rangle)\mathbb{I}[|\langle \theta^*,\, x\rangle| \leq T]]\right|.$$

Suppose that we can prove that

$$\mathbb{E}\Big[\widetilde{\varphi}_\tau(\langle \Delta,\, x\rangle)\mathbb{I}[|\langle \theta^*,\, x\rangle| \leq T]\Big] \geq \frac{3}{4}\alpha \tag{9.99a}$$

and

$$\mathbb{P}\Big[Z_n(r) > \alpha/2 + c_4 r\mu_n(\Phi^*)\Big] \leq \exp\left(-c_2\frac{nr^2\mu_n^2(\Phi^*)}{\sigma^2} - c_2 n\right). \tag{9.99b}$$

The bound (9.98) with $c_3 = \alpha/4$ then follows for all vectors with unit Euclidean norm and $\Phi(\Delta) \leq r$. Accordingly, we prove the bounds (9.99a) and (9.99b) here for a fixed radius r. A peeling argument can be used to extend it to all radii, as in the proof of Theorem 9.34, with the probability still upper bounded by $c_1 e^{-c_2 n}$.

Proof of the expectation bound (9.99a)*:* We claim that it suffices to show that

$$\mathbb{E}\Big[\widetilde{\varphi}_\tau(\langle \Delta,\, x\rangle)\Big] \overset{\text{(i)}}{\geq} \frac{7}{8}\alpha, \quad \text{and} \quad \mathbb{E}\Big[\widetilde{\varphi}_\tau(\langle \Delta,\, x\rangle)\ \mathbb{I}[|\langle \theta^*,\, x\rangle| > T]\Big] \overset{\text{(ii)}}{\leq} \frac{1}{8}\alpha.$$

Indeed, if these two inequalities hold, then we have

$$\begin{aligned}\mathbb{E}[\widetilde{\varphi}_\tau(\langle \Delta,\, x\rangle)\mathbb{I}[|\langle \theta^*,\, x\rangle| \leq T]] &= \mathbb{E}[\widetilde{\varphi}_\tau(\langle \Delta,\, x\rangle)] - \mathbb{E}[\widetilde{\varphi}_\tau(\langle \Delta,\, x\rangle)\mathbb{I}[|\langle \theta^*,\, x\rangle| > T]]\\ &\geq \left\{\frac{7}{8} - \frac{1}{8}\right\}\alpha = \frac{3}{4}\alpha.\end{aligned}$$

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

We now prove inequalities (i) and (ii). Beginning with inequality (i), we have

$$\begin{aligned}\mathbb{E}[\widetilde{\varphi}_\tau(\langle \Delta,\, x\rangle)] \geq \mathbb{E}\Big[\,\langle \Delta,\, x\rangle^2\, \mathbb{I}[|\langle \Delta,\, x\rangle| \leq \tau]\Big] &= \mathbb{E}[\langle \Delta,\, x\rangle^2] - \mathbb{E}\Big[\,\langle \Delta,\, x\rangle^2\, \mathbb{I}[|\langle \Delta,\, x\rangle| > \tau]\Big] \\ &\geq \alpha - \mathbb{E}\Big[\,\langle \Delta,\, x\rangle^2\, \mathbb{I}[|\langle \Delta,\, x\rangle| > \tau]\Big],\end{aligned}$$

so that it suffices to show that the last term is at most $\alpha/8$. By the condition (9.94) and Markov's inequality, we have

$$\mathbb{P}[|\langle \Delta,\, x\rangle| > \tau] \leq \frac{\mathbb{E}[\langle \Delta,\, x\rangle^4]}{\tau^4} \leq \frac{\beta}{\tau^4}$$

and

$$\mathbb{E}[\langle \Delta,\, x\rangle^4] \leq \beta.$$

Recalling that $\tau = K$ when $\delta = 1$, applying the Cauchy–Schwarz inequality yields

$$\mathbb{E}\Big[\,\langle \Delta,\, x\rangle^2\, \mathbb{I}[|\langle \Delta,\, x\rangle| > \tau]\Big] \leq \sqrt{\mathbb{E}[\langle \Delta,\, x\rangle^4]}\ \sqrt{\mathbb{P}[|\langle \Delta,\, x\rangle| > \tau]}\ \leq \frac{\beta}{K^2},$$

so that setting $K^2 = 8\beta/\alpha$ guarantees an upper bound of $\alpha/8$, which in turn implies inequality (i) by our earlier reasoning.

Turning to inequality (ii), since

$$\widetilde{\varphi}_\tau(\langle \Delta,\, x\rangle) \leq \langle \Delta,\, x\rangle^2 \quad \text{and} \quad \mathbb{P}[|\langle \theta^*,\, x\rangle| \geq T] \leq \frac{\beta \|\theta^*\|_2^4}{T^4},$$

the Cauchy–Schwarz inequality implies that

$$\mathbb{E}[\widetilde{\varphi}_\tau(\langle \Delta,\, x\rangle)\mathbb{I}[\,|\,\langle \theta^*,\, x\rangle\,|\, > T]] \leq \frac{\beta \|\theta^*\|_2^2}{T^2}.$$

Thus, setting $T^2 = 8\beta\|\theta^*\|_2^2/\alpha$ guarantees inequality (ii).

Proof of the tail bound (9.99b)*:* By our choice $\tau = K$, the empirical process defining $Z_n(r)$ is based on functions bounded in absolute value by K^2. Thus, the functional Hoeffding inequality (Theorem 3.26) implies that

$$\mathbb{P}[Z_n(r) \geq \mathbb{E}[Z_n(r)] + r\mu_n(\Phi^*) + \alpha/2] \leq e^{-c_2 n r^2 \mu_n^2(\Phi^*) - c_2 n}.$$

As for the expectation, letting $\{\varepsilon_i\}_{i=1}^n$ denote an i.i.d. sequence of Rademacher variables, the usual symmetrization argument (Proposition 4.11) implies that

$$\mathbb{E}[Z_n(r)] \leq 2 \sup \mathbb{E}_{x,\varepsilon}\Bigg[\sup_{\substack{\|\Delta\|_2 = 1 \\ \Phi(\Delta) \leq r}} \Bigg|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \widetilde{\varphi}_\tau(\langle \Delta,\, x_i\rangle)\ \mathbb{I}[\,|\,\langle \theta^*,\, x_i\rangle\,|\, \leq T]\Bigg|\,\Bigg].$$

Since $\mathbb{I}[\,|\,\langle \theta^*,\, x_i\rangle\,|\, \leq T] \leq 1$ and $\widetilde{\varphi}_\tau$ is Lipschitz with parameter $2K$, the contraction principle yields

$$\mathbb{E}[Z_n(r)] \leq 8K\, \mathbb{E}_{x,\varepsilon}\Bigg[\sup_{\substack{\|\Delta\|_2 = 1 \\ \Phi(\Delta) \leq r}} \Bigg|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \,\langle \Delta,\, x_i\rangle\Bigg|\,\Bigg] \leq 8Kr\, \mathbb{E}\Big[\Phi^*\Big(\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i\Big)\Big],$$

where the final step follows by applying Hölder's inequality using Φ and its dual Φ^*. □

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

9.9 Appendix: Star-shaped property

Recall the set $\mathbb{C}$ previously defined in Proposition 9.13. In this appendix, we prove that $\mathbb{C}$ is star-shaped around the origin, meaning that if $\Delta \in \mathbb{C}$, then $t\Delta \in \mathbb{C}$ for all $t \in [0, 1]$. This property is immediate whenever $\theta^* \in \mathbb{M}$, since $\mathbb{C}$ is then a cone, as illustrated in Figure 9.7(a). Now consider the general case, when $\theta^* \notin \mathbb{M}$. We first observe that for any $t \in (0, 1]$,

$$\Pi_{\bar{\mathbb{M}}}(t\Delta) = \arg\min_{\theta \in \bar{\mathbb{M}}} \|t\Delta - \theta\| = t \;\arg\min_{\theta \in \bar{\mathbb{M}}} \left\|\Delta - \frac{\theta}{t}\right\| = t\,\Pi_{\bar{\mathbb{M}}}(\Delta),$$

using the fact that θ/t also belongs to the subspace $\bar{\mathbb{M}}$. A similar argument can be used to establish the equality $\Pi_{\bar{\mathbb{M}}^\perp}(t\Delta) = t\Pi_{\bar{\mathbb{M}}^\perp}(\Delta)$. Consequently, for all $\Delta \in \mathbb{C}$, we have

$$\begin{aligned}\Phi(\Pi_{\bar{\mathbb{M}}^\perp}(t\Delta)) = \Phi(t\Pi_{\bar{\mathbb{M}}^\perp}(\Delta)) &\overset{\text{(i)}}{=} t\,\Phi(\Pi_{\bar{\mathbb{M}}^\perp}(\Delta)) \\ &\overset{\text{(ii)}}{\leq} t\,\left\{3\,\Phi(\Pi_{\bar{\mathbb{M}}}(\Delta)) + 4\Phi(\theta^*_{\mathbb{M}^\perp})\right\},\end{aligned}$$

where step (i) uses the fact that any norm is positive homogeneous,[4] and step (ii) uses the inclusion $\Delta \in \mathbb{C}$. We now observe that $3\,t\,\Phi(\Pi_{\bar{\mathbb{M}}}(\Delta)) = 3\,\Phi(\Pi_{\bar{\mathbb{M}}}(t\Delta))$, and moreover, since $t \in (0, 1]$, we have $4t\,\Phi(\theta^*_{\mathbb{M}^\perp}) \leq 4\Phi(\theta^*_{\mathbb{M}^\perp})$. Putting together the pieces, we find that

$$\Phi(\Pi_{\bar{\mathbb{M}}^\perp}(t\Delta)) \leq \;3\,\Phi(\Pi_{\bar{\mathbb{M}}}(t\Delta)) + 4\,t\Phi(\theta^*_{\mathbb{M}^\perp}) \;\leq 3\,\Phi(\Pi_{\bar{\mathbb{M}}}(t\Delta)) + 4\Phi(\theta^*_{\mathbb{M}^\perp}),$$

showing that $t\Delta \in \mathbb{C}$ for all $t \in (0, 1]$, as claimed.

9.10 Bibliographic details and background

The definitions of decomposable regularizers and restricted strong convexity were introduced by Negahban et al. (2012), who first proved a version of Theorem 9.19. Restricted strong convexity is the natural generalization of a restricted eigenvalue to the setting of general (potentially non-quadratic) cost functions, and general decomposable regularizers. A version of Theorem 9.36 was proved in the technical report (Negahban et al., 2010) for the ℓ_1-norm; note that this result allows for the second derivative ψ'' to be unbounded, as in the Poisson case. The class of decomposable regularizers includes the atomic norms studied by Chandrasekaran et al. (2012a), whereas van de Geer (2014) introduced a generalization known as weakly decomposable regularizers.

The argument used in the proof of Theorem 9.19 exploits ideas from Ortega and Rheinboldt (2000) as well as Rothman et al. (2008), who first derived Frobenius norm error bounds on the graphical Lasso (9.12). See Chapter 11 for a more detailed discussion of the graphical Lasso, and related problems concerning graphical models. The choice of regularizer defining the "good" event $\mathbb{G}(\lambda_n)$ in Proposition 9.13 is known as the *dual norm bound*. It is a cleanly stated and generally applicable choice, sharp for many (but not all) problems. See Exercise 7.15 as well as Chapter 13 for a discussion of instances in which it can be improved. These types of dual-based quantities also arise in analyses of exact recovery based on random projections; see the papers by Mendelson et al. (2007) and Chandrasekaran et al. (2012a) for geometric perspectives of this type.

The ℓ_1/ℓ_2 group Lasso norm from Example 9.3 was introduced by Yuan and Lin (2006);

[4] Explicitly, for any norm and non-negative scalar t, we have $\|tx\| = t\|x\|$.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

see also Kim et al. (2006). As a convex program, it is a special case of second-order cone program (SOCP), for which there are various efficient algorithms (Bach et al., 2012; Boyd and Vandenberghe, 2004). Turlach et al. (2005) studied the ℓ_1/ℓ_∞ version of the group Lasso norm. Several groups (Zhao et al., 2009; Baraniuk et al., 2010) have proposed unifying frameworks that include these group-structured norms as particular cases. See Bach et al. (2012) for discussion of algorithmic issues associated with optimization involving group sparse penalties. Jacob et al. (2009) introduced the overlapping group Lasso norm discussed in Example 9.4, and provide detailed discussion of why the standard group Lasso norm with overlap fails to select unions of groups. A number of authors have investigated the statistical benefits of the group Lasso versus the ordinary Lasso when the underlying regression vector is group-sparse; for instance, Obozinski et al. (2011) study the problem of variable selection, whereas the papers (Baraniuk et al., 2010; Huang and Zhang, 2010; Lounici et al., 2011) provide guarantees on the estimation error. Negahban and Wainwright (2011a) study the variable selection properties of ℓ_1/ℓ_∞-regularization for multivariate regression, and show that, while it can be more statistically efficient than ℓ_1-regularization with complete shared overlap, this gain is surprisingly non-robust: it is very easy to construct examples in which it is outperformed by the ordinary Lasso. Motivated by this deficiency, Jalali et al. (2010) study a decomposition-based estimator, in which the multivariate regression matrix is decomposed as the sum of an elementwise-sparse and row-sparse matrix (as in Section 9.7), and show that it adapts in the optimal way. The adaptive guarantee given in Corollary 9.31 is of a similar flavor, but as applied to the estimation error as opposed to variable selection.

Convex relaxations based on nuclear norm introduced in Example 9.8 have been the focus of considerable research; see Chapter 10 for an in-depth discussion.

The Φ^*-norm restricted curvature conditions discussed in Section 9.3 are a generalization of the notion of ℓ_∞-restricted eigenvalues (van de Geer and Bühlmann, 2009; Ye and Zhang, 2010; Bühlmann and van de Geer, 2011). See Exercises 7.13, 7.14 and 9.11 for some analysis of these ℓ_∞-RE conditions for the usual Lasso, and Exercise 9.14 for some analysis for Lipschitz cost functions. Section 10.2.3 provides various applications of this condition to nuclear norm regularization.

9.11 Exercises

Exercise 9.1 (Overlapping group Lasso) Show that the overlap group Lasso, as defined by the variational representation (9.10), is a valid norm.

Exercise 9.2 (Subspace projection operator) Recall the definition (9.20) of the subspace projection operator. Compute an explicit form for the following subspaces:

(a) For a fixed subset $S \subseteq \{1, 2, \ldots, d\}$, the subspace of vectors

$$\mathbb{M}(S) := \{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \notin S\}.$$

(b) For a given pair of r-dimensional subspaces $\mathbb{U}$ and $\mathbb{V}$, the subspace of matrices

$$\mathbb{M}(\mathbb{U}, \mathbb{V}) := \{\mathbf{\Theta} \in \mathbb{R}^{d \times d} \mid \text{rowspan}(\mathbf{\Theta}) \subseteq \mathbb{U},\ \text{colspan}(\mathbf{\Theta}) \subseteq \mathbb{V}\},$$

where rowspan($\mathbf{\Theta}$) and colspan($\mathbf{\Theta}$) denote the row and column spans of $\mathbf{\Theta}$.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

Exercise 9.3 (Generalized linear models) This exercise treats various cases of the generalized linear model.

(a) Suppose that we observe samples of the form $y = \langle x, \theta \rangle + w$, where $w \sim N(0, \sigma^2)$. Show that the conditional distribution of y given x is of the form (9.5) with $c(\sigma) = \sigma^2$ and $\psi(t) = t^2/2$.

(b) Suppose that y is (conditionally) Poisson with mean $\lambda = e^{\langle x, \theta \rangle}$. Show that this is a special case of the log-linear model (9.5) with $c(\sigma) \equiv 1$ and $\psi(t) = e^t$.

Exercise 9.4 (Dual norms) In this exercise, we study various forms of dual norms.

(a) Show that the dual norm of the ℓ_1-norm is the ℓ_∞-norm.

(b) Consider the general group Lasso norm

$$\Phi(u) = \|u\|_{1,\mathcal{G}(p)} = \sum_{g \in \mathcal{G}} \|u_g\|_p,$$

where $p \in [1, \infty]$ is arbitrary, and the groups are non-overlapping. Show that its dual norm takes the form

$$\Phi^*(v) = \|v\|_{\infty,\mathcal{G}(q)} = \max_{g \in \mathcal{G}} \|v_g\|_q,$$

where $q = \frac{p}{p-1}$ is the conjugate exponent to p.

(c) Show that the dual norm of the nuclear norm is the ℓ_2-operator norm

$$\Phi^*(\mathbf{N}) = |\!|\!|\mathbf{N}|\!|\!|_2 := \sup_{\|z\|_2 = 1} \|\mathbf{N}z\|_2.$$

(*Hint:* Try to reduce the problem to a version of part (a).)

Exercise 9.5 (Overlapping group norm and duality) Let $p \in [1, \infty]$, and recall the overlapping group norm (9.10).

(a) Show that it has the equivalent representation

$$\Phi(u) = \max_{v \in \mathbb{R}^d} \langle v, u \rangle \quad \text{such that } \|v_g\|_q \leq 1 \text{ for all } g \in \mathcal{G},$$

where $q = \frac{p}{p-1}$ is the dual exponent.

(b) Use part (a) to show that its dual norm is given by

$$\Phi^*(v) = \max_{g \in \mathcal{G}} \|v_g\|_q.$$

Exercise 9.6 (Boundedness of subgradients in the dual norm) Let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be a norm, and $\theta \in \mathbb{R}^d$ be arbitrary. For any $z \in \partial\Phi(\theta)$, show that $\Phi^*(z) \leq 1$.

Exercise 9.7 (Hölder's inequality) Let $\Phi : \mathbb{R}^d \to \mathbb{R}_+$ be a norm, and let $\Phi^* : \mathbb{R}^d \to \mathbb{R}_+$ be its dual norm.

(a) Show that $\big|\langle u, v \rangle\big| \leq \Phi(u)\, \Phi^*(v)$ for all $u, v \in \mathbb{R}^d$.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

(b) Use part (a) to prove Hölder's inequality for ℓ_p-norms, namely

$$\big|\langle u,\, v\rangle\big| \le \|u\|_p\, \|v\|_q,$$

where the exponents (p, q) satisfy the conjugate relation $1/p + 1/q = 1$.

(c) Let $\mathbf{Q} \succ 0$ be a positive definite symmetric matrix. Use part (a) to show that

$$\big|\langle u,\, v\rangle\big| \le \sqrt{u^{\mathrm{T}}\mathbf{Q}u}\,\sqrt{v^{\mathrm{T}}\mathbf{Q}^{-1}v} \quad \text{for all } u, v \in \mathbb{R}^d.$$

Exercise 9.8 (Complexity parameters) This exercise concerns the complexity parameter $\mu_n(\Phi^*)$ previously defined in equation (9.41). Suppose throughout that the covariates $\{x_i\}_{i=1}^n$ are drawn i.i.d., each sub-Gaussian with parameter σ.

(a) Consider the group Lasso norm (9.9) with group set $\mathcal{G}$ and maximum group size m. Show that

$$\mu_n(\Phi^*) \precsim \sigma\sqrt{\frac{m}{n}} + \sigma\sqrt{\frac{\log|\mathcal{G}|}{n}}.$$

(b) For the nuclear norm on the space of $d_1 \times d_2$ matrices, show that

$$\mu_n(\Phi^*) \precsim \sigma\sqrt{\frac{d_1}{n}} + \sigma\sqrt{\frac{d_2}{n}}.$$

Exercise 9.9 (Equivalent forms of strong convexity) Suppose that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is κ-strongly convex in the sense that

$$f(y) \ge f(x) + \langle \nabla f(x),\, y - x\rangle + \frac{\kappa}{2}\|y - x\|_2^2 \qquad \text{for all } x, y \in \mathbb{R}^d. \tag{9.100a}$$

Show that

$$\langle \nabla f(y) - \nabla f(x),\, y - x\rangle \ge \kappa\|y - x\|_2^2 \qquad \text{for all } x, y \in \mathbb{R}^d. \tag{9.100b}$$

Exercise 9.10 (Implications of local strong convexity) Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is a twice differentiable, convex function that is *locally* κ-strongly convex around x, in the sense that the lower bound (9.100a) holds for all vectors z in the ball $\mathbb{B}_2(x) := \{z \in \mathbb{R}^d \mid \|z - x\|_2 \le 1\}$. Show that

$$\langle \nabla f(y) - \nabla f(x),\, y - x\rangle \ge \kappa\|y - x\|_2 \quad \text{for all } y \in \mathbb{R}^d\backslash\mathbb{B}_2(x).$$

Exercise 9.11 (ℓ_∞-curvature and RE conditions) In this exercise, we explore the link between the ℓ_∞-curvature condition (9.56) and the ℓ_∞-RE condition (9.57). Suppose that the bound (9.56) holds with $\tau_n = c_1\sqrt{\frac{\log d}{n}}$. Show that the bound (9.57) holds with $\kappa' = \frac{\kappa}{2}$ as long as $n > c_2|S|^2 \log d$ with $c_2 = \frac{4c_1^2(1+\alpha)^4}{\kappa^2}$.

Exercise 9.12 (ℓ_1-regularization and soft thresholding) Given observations from the linear model $y = \mathbf{X}\theta^* + w$, consider the M-estimator

$$\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}^d}\left\{\frac{1}{2}\|\theta\|_2^2 - \left\langle \theta,\, \frac{1}{n}\mathbf{X}^{\mathrm{T}}y\right\rangle + \lambda_n\|\theta\|_1\right\}.$$

(a) Show that the optimal solution is always unique, and given by $\widehat{\theta} = T_{\lambda_n}(\frac{1}{n}\mathbf{X}^{\mathrm{T}}y)$, where the soft-thresholding operator T_{λ_n} was previously defined (7.6b).

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

(b) Now suppose that θ^* is s-sparse. Show that if

$$\lambda_n \geq 2\left\{\left\|\left(\frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{n} - \mathbf{I}_d\right)\theta^*\right\|_\infty + \left\|\frac{\mathbf{X}^{\mathrm{T}}w}{n}\right\|_\infty\right\},$$

then the optimal solution satisfies the bound $\|\widehat{\theta} - \theta^*\|_2 \leq \frac{3}{2}\sqrt{s}\lambda_n$.

(c) Now suppose that the covariates $\{x_i\}_{i=1}^n$ are drawn i.i.d. from a zero-mean ν-sub-Gaussian ensemble with covariance $\mathrm{cov}(x_i) = \mathbf{I}_d$, and the noise vector w is bounded as $\|w\|_2 \leq b\sqrt{n}$ for some $b > 0$. Show that with an appropriate choice of λ_n, we have

$$\|\widehat{\theta} - \theta^*\|_2 \leq 3\nu\,(\nu\|\theta^*\|_2 + b)\,\sqrt{s}\left\{\sqrt{\frac{\log d}{n}} + \delta\right\}$$

with probability at least $1 - 4e^{-\frac{n\delta^2}{8}}$ for all $\delta \in (0, 1)$.

Exercise 9.13 (From ℓ_∞ to $\{\ell_1, \ell_2\}$-bounds) In the setting of Corollary 9.27, show that any optimal solution $\widehat{\theta}$ that satisfies the ℓ_∞-bound (9.65) also satisfies the following ℓ_1- and ℓ_2-error bounds

$$\|\widehat{\theta} - \theta^*\|_1 \leq \frac{24\sigma}{\kappa}s\sqrt{\frac{\log d}{n}} \quad \text{and} \quad \|\widehat{\theta} - \theta^*\|_2 \leq \frac{12\sigma}{\kappa}\sqrt{\frac{s\log d}{n}}.$$

(*Hint:* Proposition 9.13 is relevant here.)

Exercise 9.14 (ℓ_∞-curvature for Lipschitz cost functions) In the setting of regression-type data $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$, consider a cost function whose gradient is elementwise L-Lipschitz: i.e., for any sample z and pair $\theta, \widetilde{\theta}$, the jth partial derivative satisfies

$$\left|\frac{\partial\mathcal{L}(\theta; z_i)}{\theta_j} - \frac{\partial\mathcal{L}(\widetilde{\theta}; z_i)}{\theta_j}\right| \leq L\left|x_{ij}\left\langle x_i,\, \theta - \widetilde{\theta}\right\rangle\right|. \tag{9.101}$$

The goal of this exercise is to show that such a function satisfies an ℓ_∞-curvature condition similar to equation (9.64), as required for applying Corollary 9.27.

(a) Show that for any GLM whose cumulant function has a uniformly bounded second derivative ($\|\psi''\|_\infty \leq B^2$), the elementwise Lipschitz condition (9.101) is satisfied with $L = \frac{B^2}{2}$.

(b) For a given radius $r > 0$ and ratio $\rho > 0$, define the set

$$\mathbb{T}(R; \rho) := \left\{\Delta \in \mathbb{R}^d \mid \frac{\|\Delta\|_1}{\|\Delta\|_\infty} \leq \rho, \text{ and } \|\Delta\|_\infty \leq r\right\},$$

and consider the random vector $V \in \mathbb{R}^d$ with elements

$$V_j := \frac{1}{4Lr\rho}\sup_{\Delta \in \mathbb{T}(r;\rho)}\left|\frac{1}{n}\sum_{i=1}^n f_j(\Delta; z_i)\right|, \quad \text{for } j = 1, \ldots, d,$$

where, for each fixed vector Δ,

$$f_j(\Delta; z_i) := \left\{\frac{\partial\mathcal{L}(\theta^* + \Delta; z_i)}{\theta_j} - \frac{\partial\mathcal{L}(\theta^*; z_i)}{\theta_j}\right\} - \left\{\frac{\partial\bar{\mathcal{L}}(\theta^* + \Delta)}{\theta_j} - \frac{\partial\bar{\mathcal{L}}(\theta^*)}{\theta_j}\right\}$$

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press

is a zero-mean random variable. For each $\lambda > 0$, show that

$$\mathbb{E}_x[e^{\lambda \|V\|_\infty}] \leq d\, \mathbb{E}_{x,\varepsilon}\left[\exp\left(\lambda \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i x_i x_i^{\mathrm{T}} \right\|_\infty\right)\right].$$

(c) Suppose that the covariates $\{x_i\}_{i=1}^n$ are sampled independently, with each x_{ij} following a zero-mean σ-sub-Gaussian distribution. Show that for all $t \in (0, \sigma^2)$,

$$\mathbb{P}[\|V\|_\infty \geq t] \leq 2d^2 e^{-\frac{nt^2}{2\sigma^4}}.$$

(d) Suppose that the population function $\bar{\mathcal{L}}$ satisfies the ℓ_∞- curvature condition

$$\|\nabla \bar{\mathcal{L}}(\theta^* + \Delta) - \nabla \bar{\mathcal{L}}(\theta^*)\|_\infty \geq \kappa \,\|\Delta\|_\infty \quad \text{for all } \Delta \in \mathbb{T}(r; \rho).$$

Use this condition and the preceding parts to show that

$$\|\nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \mathcal{L}_n(\theta^*)\|_\infty \geq \kappa \|\Delta\|_\infty - 16\, L\, \sigma^2 \sqrt{\frac{\log d}{n}} \rho\, r \qquad \text{for all } \Delta \in \mathbb{T}(r; \rho)$$

with probability at least $1 - e^{-4 \log d}$.

https://doi.org/10.1017/9781108627771.009 Published online by Cambridge University Press