Bounds on Statistical Error for Regularized Loss Minimization

Recall: Standard linear model

$$y = X\theta^* + w \quad , \quad y, w \in \mathbb{R}^n; \; X \in \mathbb{R}^{n \times d}; \; \theta^* \in \mathbb{R}^d$$

Previously, we considered the setting where $n >> d$. Then, we have

$$\|X\Delta\|_2^2 \geq \lambda_{\min}(X^TX) \cdot \|\Delta\|_2^2 \qquad \forall \Delta \in \mathbb{R}^d \qquad -(*)$$

and $\lambda_{\min}(X^TX) > 0$ whenever X has full column rank. In particular, by taking $\Delta = \hat{\theta} - \theta^*$, where $\hat{\theta}$ is the least squares estmeator, the eigenvalue condition $(*)$ provides a first step in bounding the statistical error $\|\Delta\|_2$.

Now, in contemporary applications, we often have $n << d$. In such setting, we need additional constraints on the model in order to be able to estimate θ^*. Typically, it is assumed that θ^* has some low-dimensional structure, such as sparsity, low-rankness, etc. In this case, a natural family of estimators is given by

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \mathcal{L}(\theta, \{Z_i\}_{i=1}^n) + \lambda R(\theta) \right\}, \qquad -(\dagger)$$

where $Z_1, Z_2, \ldots, Z_n$ are observations drawn from a space $\mathcal{Z}$; $\mathcal{L} : \mathbb{R}^d \times \mathcal{Z}^n \to \mathbb{R}$ is a smooth convex loss function; $R : \mathbb{R}^d \to \mathbb{R}_+$ is a norm regularizer that aims to induce the desired low-dimensional structure in $\hat{\theta}$; $\lambda > 0$ is a user-defined regularization parameter.

For example, in the standard linear model, we have

$$z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R},$$

where x_i^T is the i^{th} row of X;

$$\mathcal{L}(\theta, \{z_i\}_{i=1}^n) = \frac{1}{2}\|y - X\theta\|_2^2.$$

In the case where θ^* is known to be sparse, a commonly used regularizer is

$$R(\theta) = \|\theta\|_1.$$

Observe that in the $n \ll d$ setting, we always have $\lambda_{\min}(X^TX) = 0$. Thus, it is hopeless to use $(\star)$ to bound the statistical error. To proceed, we need to understand how the regularizer penalizes the deviation from the low-dimensional structure it aims to induce. As it turns out, a key tool for this purpose is the notion of decomposability.

Decomposability of the Regularizer

Let $\mathcal{M} \subseteq \bar{\mathcal{M}} \subseteq \mathbb{R}^d$ be a pair of subspaces.

- $\mathcal{M}$ is the model subspace and intends to capture the constraints specified by the model.
- $\bar{\mathcal{M}}^\perp = \{ v \in \mathbb{R}^d : u^Tv = 0 \ \ \forall u \in \bar{\mathcal{M}} \}$ is the perturbation subspace and captures deviations from the model subspace.
- For simplicity, one can take $\mathcal{M} = \bar{\mathcal{M}}$, but the case where $\mathcal{M} \subsetneq \bar{\mathcal{M}}$ can be useful in some settings.

• Definition: The regularizer R is decomposable wrt $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$ if

$$R(\theta+\gamma) = R(\theta) + R(\gamma) \quad \forall \theta \in \mathcal{M}, \gamma \in \bar{\mathcal{M}}^\perp. \quad \text{—(D)}$$

To motivate this definition, observe that since R is a norm, it satisfies $R(\theta+\gamma) \leq R(\theta) + R(\gamma)$ (triangle inequality). Since γ belongs to the perturbation subspace $\bar{\mathcal{M}}^\perp$ and represents deviation from the model subspace $\mathcal{M}$, the regularizer R should penalize it in a maximal manner. This gives rise to (D).

Note that in the definition, we have the flexibility to choose the pair $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$. In general, there are many such pairs that can result in the decomposability of R. Take, for instance, $\mathcal{M} = \mathbb{R}^d$ and $\bar{\mathcal{M}}^\perp = \{0\}$. In order to have an effective decomposition, however, it is desirable to have

(i) $\theta^*_{\mathcal{M}} \triangleq \Pi_{\mathcal{M}}(\theta^*) \cong \theta^*$, so that the model subspace $\mathcal{M}$ closely captures the constraints on the ground-truth θ^*;

(ii) $\mathcal{M}$ remains relatively small, so that elements that are inconsistent with the model will fall into the perturbation subspace $\bar{\mathcal{M}}^\perp$ and get maximally penalized by R due to decomposability.

Let us now take a look at two examples.

Example 1: Sparse vectors and l_1-regularization

- model: s-sparse vectors in $\mathbb{R}^d$; $R(\theta) = \|\theta\|_1$
- model subspace: For any $S \subseteq \{1, \cdots, d\}$ of cardinality s, define

$$\mathcal{M} \triangleq \mathcal{M}(S) = \{\theta \in \mathbb{R}^d : \theta_j = 0 \ \forall j \notin S\}$$

Note that if θ^* is supported on S, then $\Pi_{\mathcal{M}}(\theta^*) = \theta^*$.

- perturbation subspace: Take $\bar{\mathcal{M}}(S) = \mathcal{M}(S)$, so that

$$\bar{\mathcal{M}}^{\perp} \triangleq \bar{\mathcal{M}}(S)^{\perp} = \mathcal{M}(S)^{\perp}$$

Now, the decomposability of R wrt $(\mathcal{M}(S), \bar{\mathcal{M}}(S)^{\perp})$ follows by observing that

$$\mathcal{M}(S)^{\perp} = \{\theta \in \mathbb{R}^d : \theta_j = 0 \ \forall j \in S\} \quad \text{(check)}$$

Example 2: Group-sparse vectors and mixed norm regularization

- Motivation: groups of coefficients likely to be zero or non-zero simultaneously
- model: partition $\{1, \cdots, d\}$ into N disjoint groups $G_1, \cdots, G_N$; number of selected groups, denoted by s, should be small;

$$R(\theta) = \sum_{i=1}^{N} \|\theta_{G_i}\|_p, \quad p \in [1, \infty] \quad \text{(mixed } \ell_{1,p}\text{-norm)}$$

- model subspace: For any $S \subseteq \{1, \cdots, N\}$ of cardinality s, define

$$\mathcal{M} \triangleq \mathcal{M}(S) = \{\theta \in \mathbb{R}^d : \theta_{G_i} = 0 \ \forall i \in S\}$$

- perturbation subspace: Take $\bar{\mathcal{M}}(S) = \mathcal{M}(S)$, so that

$$\bar{\mathcal{M}}^{\perp} \triangleq \bar{\mathcal{M}}(S)^{\perp} = \mathcal{M}(S)^{\perp}$$

Again, the decomposability of R wrt $(\mathcal{M}(S), \bar{\mathcal{M}}(S)^{\perp})$ follows by observing that

$$\mathcal{M}(S)^{\perp} = \{\theta \in \mathbb{R}^d : \theta_{G_i} = 0 \ \forall i \in S\}$$

Now, let us turn to a key consequence of decomposability, which allows us to localize the error vector $\hat{\Delta} = \hat{\theta} - \theta^*$.

Proposition 1: Suppose that $\mathcal{L}$ is smooth and convex, $\lambda \geq 2R^*(\nabla \mathcal{L}(\theta^*))$, and R is decomposable wrt $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$. Then,

$$\hat{\Delta} \in \mathcal{C} \triangleq \{\Delta \in \mathbb{R}^d : R(\Delta_{\bar{\mathcal{M}}^{\perp}}) \leq 3R(\Delta_{\bar{\mathcal{M}}}) + 4R(\theta^*_{\mathcal{M}^{\perp}})\}$$

Observe that if $\theta^* \in \mathcal{M}$, then $R(\theta^*_{\mathcal{M}^\perp}) = 0$. In this case, it can be easily verified $\mathcal{C}$ is a cone; i.e., if $\Delta \in \mathcal{C}$, then $t\Delta \in \mathcal{C}$ $\forall t > 0$. Intuitively, the term $R(\theta^*_{\mathcal{M}^\perp})$ accounts for the misspecification of the model.

- Proof of Proposition 1

  - Define
$$D(\Delta) = \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda\left(R(\theta^* + \Delta) - R(\theta^*)\right).$$
By the optimality of $\hat{\theta}$ for (†), we have $D(\hat{\Delta}) \leq 0$, where $\hat{\Delta} = \hat{\theta} - \theta^*$.

  - Claim 1: $R(\theta^* + \Delta) - R(\theta^*) \geq R(\Delta_{\bar{\mathcal{M}}^\perp}) - R(\Delta_{\bar{\mathcal{M}}}) - 2R(\theta^*_{\mathcal{M}^\perp})$

    Claim 2: If $\lambda \geq 2R^*(\nabla\mathcal{L}(\theta^*))$ and $\mathcal{L}$ is convex, then
$$\mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) \geq -\frac{\lambda}{2}\left[R(\Delta_{\bar{\mathcal{M}}}) + R(\Delta_{\bar{\mathcal{M}}^\perp})\right].$$

  - With these two claims, we have
$$\begin{aligned} 0 \geq D(\hat{\Delta}) &\geq \lambda\left[R(\Delta_{\bar{\mathcal{M}}^\perp}) - R(\Delta_{\bar{\mathcal{M}}}) - 2R(\theta^*_{\mathcal{M}^\perp})\right] \\ &\quad - \frac{\lambda}{2}\left[R(\Delta_{\bar{\mathcal{M}}}) + R(\Delta_{\bar{\mathcal{M}}^\perp})\right] \\ &= \frac{\lambda}{2}\left[R(\Delta_{\bar{\mathcal{M}}^\perp}) - 3R(\Delta_{\bar{\mathcal{M}}}) - 4R(\theta^*_{\mathcal{M}^\perp})\right]. \end{aligned}$$

  - Proof of Claim 1:

    We compute $R(\theta^* + \Delta) = R(\theta^*_{\mathcal{M}} + \theta^*_{\mathcal{M}^\perp} + \Delta_{\bar{\mathcal{M}}} + \Delta_{\bar{\mathcal{M}}^\perp})$
$$\begin{aligned} &\geq R(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) - R(\theta^*_{\mathcal{M}^\perp} + \Delta_{\bar{\mathcal{M}}}) \\ &\geq R(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) - R(\theta^*_{\mathcal{M}^\perp}) - R(\Delta_{\bar{\mathcal{M}}}) \quad \text{(triangle inequality)} \\ &= R(\theta^*_{\mathcal{M}}) + R(\Delta_{\bar{\mathcal{M}}^\perp}) - R(\theta^*_{\mathcal{M}^\perp}) - R(\Delta_{\bar{\mathcal{M}}}) \quad \text{(decomposability of } R\text{)} \end{aligned}$$
On the other hand,
$$R(\theta^*) \leq R(\theta^*_{\mathcal{M}}) + R(\theta^*_{\mathcal{M}^\perp}).$$

It follows that

$$R(\theta^* + \Delta) - R(\theta^*) \geq R(\Delta_{\bar{m}^\perp}) - R(\Delta_{\bar{m}}) - 2R(\theta^*_{m^\perp}).$$

□

- Proof of Claim 2

By convexity of $\mathcal{L}$,

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

Using the generalized Cauchy-Schwarz and triangle inequalities, we have

$$|\nabla\mathcal{L}(\theta^*)^T\Delta| \leq R^*(\nabla\mathcal{L}(\theta^*)) \cdot R(\Delta)$$

$$\leq \frac{\lambda}{2}\left[R(\Delta_{\bar{m}}) + R(\Delta_{\bar{m}^\perp})\right].$$

The desired result follows. □

Now, to obtain a bound on the statistical error, it suffices to show that $\mathcal{L}$ is not "too flat" on the set $\mathcal{C}$. Motivated by our earlier development, we formalize this using the following notion of strong convexity:

Definition (Restricted Strong Convexity (RSC))

We say that $\mathcal{L}$ satisfies the RSC property if there exist a constant $\kappa > 0$ and a function $\tau(\cdot)$ such that

$$\mathcal{L}(\theta^* + \Delta) \geq \mathcal{L}(\theta^*) + \nabla\mathcal{L}(\theta^*)^T\Delta + \kappa\|\Delta\|_2^2 - \tau^2(\theta^*) \quad \forall \Delta \in \mathcal{C}.$$

The above definition gives a restricted notion of strong convexity in the sense that (i) it holds only at the point θ^*, (ii) it holds only for those directions $\Delta \in \mathcal{C}$, and (iii) it allows a tolerance function $\tau(\cdot)$. We shall see the role of $\tau(\cdot)$ in some applications later. For now, one can take $\tau(\cdot) \equiv 0$ for simplicity.

Armed with the above definition, we have the following statistical error bound.

Theorem 1: Under the assumptions of Proposition 1 and the additional assumption that $\mathcal{L}$ is RSC, we have

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \frac{9\lambda^2}{4\kappa^2}\Psi^2(\bar{M}) + \frac{2}{\kappa}\left[\tau^2(\theta^*) + 2\lambda R(\theta^*_{M^\perp})\right], \quad -(\dagger\dagger)$$

where

$$\Psi(M) = \sup_{u \in M \setminus \{0\}} \frac{R(u)}{\|u\|_2}$$

can be regarded as the Lipschitz constant of R over the subspace M.

Remarks:

(1) $(\dagger\dagger)$ actually yields a family of bounds that depends on the choice of $(M, \bar{M}^\perp)$.

(2) From $(\dagger\dagger)$, we see that the statistical error can be attributed to two terms (assuming $\tau(\cdot) \equiv 0$): the estimation error $\frac{9\lambda^2}{4\kappa^2}\Psi^2(\bar{M})$, and the approximation error $\frac{4\lambda}{\kappa}R(\theta^*_{M^\perp})$. Since $M \subseteq \bar{M}$ by construction, we see that the larger the M, the lower the approximation error (since $R(\theta^*_{M^\perp})$ will be smaller), but the larger the estimation error (since $\Psi(\bar{M})$ will be bigger).

(3) To obtain more concrete conclusions for specific applications, we essentially need to estimate all the parameters in the theorem.

Proof of Theorem 1: Recall that $\hat{\Delta} = \hat{\theta} - \theta^*$ and

$$\mathcal{D}(\Delta) = \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \lambda\left[R(\theta^* + \Delta) - R(\theta^*)\right].$$

Intuitively, for $\Delta \in \mathcal{C}$, we want that if $|\mathcal{D}(\Delta)|$ is small, then so is $\|\Delta\|_2$. Since $\mathcal{D}(\hat{\Delta}) \leq 0$, this motivates the following:

- Claim 3: For any $\delta > 0$, if $\mathcal{D}(\Delta) > 0$ for all $\Delta \in K(\delta) \triangleq \mathcal{C} \cap \{\Delta \in \mathbb{R}^d : \|\Delta\|_2 = \delta\}$, then $\|\hat{\Delta}\|_2 \leq \delta$.

In view of Claim 3, to prove Theorem 1, it suffices to show that $\mathcal{D}(\Delta) > 0$ over $K(\delta)$ for some suitably chosen $\delta > 0$. Towards that end, for any $\Delta \in K(\delta)$, we compute

$$\begin{aligned}
\mathcal{D}(\Delta) &\geq \nabla \mathcal{L}(\theta^*)^T \Delta + \kappa \|\Delta\|_2^2 - \tau^2(\theta^*) + \lambda \left[R(\theta^* + \Delta) - R(\theta^*) \right] && \text{(RSC)} \\
&\geq \nabla \mathcal{L}(\theta^*)^T \Delta + \kappa \|\Delta\|_2^2 - \tau^2(\theta^*) \\
&\quad + \lambda \left[R(\Delta_{\bar{\mathcal{M}}^\perp}) - R(\Delta_{\bar{\mathcal{M}}}) - 2R(\theta^*_{\mathcal{M}^\perp}) \right] && \text{(Claim 1)} \\
&\geq \kappa \|\Delta\|_2^2 - \tau^2(\theta^*) + \lambda \left[R(\Delta_{\bar{\mathcal{M}}^\perp}) - R(\Delta_{\bar{\mathcal{M}}}) - 2R(\theta^*_{\mathcal{M}^\perp}) \right] \\
&\quad - R^*(\nabla \mathcal{L}(\theta^*)) \cdot R(\Delta) && \text{(generalized Cauchy-Schwarz inequality)} \\
&\geq \kappa \|\Delta\|_2^2 - \tau^2(\theta^*) + \frac{\lambda}{2} \left[R(\Delta_{\bar{\mathcal{M}}^\perp}) - 3R(\Delta_{\bar{\mathcal{M}}}) - 4R(\theta^*_{\mathcal{M}^\perp}) \right] \\
&\qquad \text{(since } \lambda \geq 2R^*(\nabla \mathcal{L}(\theta^*)) \text{ and } R(\Delta) \leq R(\Delta_{\bar{\mathcal{M}}^\perp}) + R(\Delta_{\bar{\mathcal{M}}})\text{)} \\
&\geq \kappa \|\Delta\|_2^2 - \tau^2(\theta^*) - \frac{\lambda}{2} \left[3R(\Delta_{\bar{\mathcal{M}}}) + 4R(\theta^*_{\mathcal{M}^\perp}) \right]. \\
&\qquad \text{(since } R(\Delta_{\bar{\mathcal{M}}^\perp}) \geq 0\text{)}
\end{aligned}$$

Now, by definition of $\Psi(\bar{\mathcal{M}})$ and the non-expansiveness of projection, we have

$$R(\Delta_{\bar{\mathcal{M}}}) \leq \Psi(\bar{\mathcal{M}}) \cdot \|\Delta_{\bar{\mathcal{M}}}\|_2 = \Psi(\bar{\mathcal{M}}) \cdot \|\Pi_{\bar{\mathcal{M}}}(\Delta) - \Pi_{\bar{\mathcal{M}}}(0)\|_2 \leq \Psi(\bar{\mathcal{M}}) \cdot \|\Delta\|_2$$

It follows that

$$\mathcal{D}(\Delta) \geq \kappa \|\Delta\|_2^2 - \tau^2(\theta^*) - \frac{\lambda}{2} \left[3\Psi(\bar{\mathcal{M}}) \cdot \|\Delta\|_2 + 4R(\theta^*_{\mathcal{M}^\perp}) \right].$$

The RHS of the above inequality is quadratic in $\|\Delta\|_2$ and will be positive when

$$\|\Delta\|_2 \geq \delta = \frac{3\lambda}{4\kappa} \Psi(\bar{\mathcal{M}}) + \frac{1}{2\kappa} \sqrt{\frac{9\lambda^2}{4} \Psi^2(\bar{\mathcal{M}}) + 4\kappa \left(\tau^2(\theta^*) + 2\lambda R(\theta^*_{\mathcal{M}^\perp}) \right)}$$

Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2 \ \forall a, b \geq 0$ yields the desired result. □

Now, let us return to the proof of Claim 3. It relies on the following result:

- Sub-Claim: $\mathcal{C}$ is star-shaped; i.e., for any $\Delta \in \mathcal{C}$,

$$\{t\Delta : t \in (0,1)\} \subseteq \mathcal{C}.$$

Proof of Claim 3.

Suppose $\|\hat{\Delta}\|_2 > \delta$. Then, the line joining 0 and $\hat{\Delta}$ will intersect $K(\delta)$ at some $t^*\hat{\Delta}$, where $t^* \in (0,1)$. Since $\mathcal{D}$ is convex,

$$\mathcal{D}(t^*\hat{\Delta}) = \mathcal{D}(t^*\hat{\Delta} + (1-t^*)0) \leq t^*\mathcal{D}(\hat{\Delta}) + (1-t^*)\mathcal{D}(0)$$
$$= t^*\mathcal{D}(\hat{\Delta}) \leq 0.$$

Noting $t^*\hat{\Delta} \in K(\delta)$, this contradicts the assumption. □

Proof of Sub-Claim:

If $\theta^* \in \mathcal{M}$, observe that $\mathcal{C}$ is a cone, since $R(\theta^*_{\bar{\mathcal{M}}^\perp}) = 0$. Thus, the result holds.

Now, suppose $\theta^* \notin \mathcal{M}$. Observe for any $t \in (0,1)$

$$\Pi_{\bar{\mathcal{M}}}(t\Delta) = \underset{\gamma \in \bar{\mathcal{M}}}{\operatorname{argmin}} \|t\Delta - \gamma\|_2$$
$$= t \underset{\gamma \in \bar{\mathcal{M}}}{\operatorname{argmin}} \|\Delta - \gamma/t\|_2$$

Since $\bar{\mathcal{M}}$ is a subspace, we have $t\gamma \in \bar{\mathcal{M}}$. Thus,

$$\Pi_{\bar{\mathcal{M}}}(t\Delta) = t\,\Pi_{\bar{\mathcal{M}}}(\Delta).$$

Similarly,

$$\Pi_{\bar{\mathcal{M}}^\perp}(t\Delta) = t\,\Pi_{\bar{\mathcal{M}}^\perp}(\Delta).$$

Hence,

$$R(\Pi_{\bar{\mathcal{M}}^\perp}(t\Delta)) = R(t\,\Pi_{\bar{\mathcal{M}}^\perp}(\Delta))$$
$$= t \cdot R(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta))$$
$$\leq t\left[3R(\Delta_{\bar{\mathcal{M}}}) + 4R(\theta^*_{\bar{\mathcal{M}}^\perp})\right]$$

Since

$$t \cdot 3R(\Delta_{\bar{m}}) = 3R(t\,\Pi_{\bar{m}}(\Delta)) = 3R(\Pi_{\bar{m}}(t\Delta))$$

by the homogeneity of R and linearity of $\Pi_{\bar{m}}(\cdot)$, and

$$4t \cdot R(\theta^*_{\bar{m}^\perp}) \leq 4R(\theta^*_{\bar{m}^\perp})$$

because $t \in (0,1)$, we conclude that

$$R(\Pi_{\bar{m}^\perp}(t\Delta)) \leq 3R(\Pi_{\bar{m}}(t\Delta)) + 4R(\theta^*_{\bar{m}^\perp});$$

i.e., $t\Delta \in \mathcal{C}$ $\forall t \in (0,1)$, as desired.