Decomposability and restricted strong convexity

Shuoxun Xu

Department of Mathematics, HKUST

April 6, 2022

Introduction

- 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.

Goals

• Introduce general family of regularized M-estimators
• Develop techniques for bounding the associated estimation error for high-dimensional problems

Notations and problem setting

- Consider a index indexed family of probability distributions $\{\mathbb{P}_\theta, \theta \in \Omega\}$, where θ is the parameter to be estimated and Ω is its possible space.
- Observe n samples $Z_1^n = (Z_1, \ldots, Z_n)$, $Z_i \in \mathcal{Z}$ and is drawn i.i.d. from $\mathbb{P}$.

For the well-specified case, $\mathbb{P}$ is a member of the parameterized space, say $\mathbb{P} = \mathbb{P}_{\theta^*}$. Our goal is to estimate θ^*.

For the mis-specified models, the target parameter θ^* is defined as minimizer of the population cost function.

Notations and problem setting

- The cost function: $\mathcal{L}_n : \Omega \times \mathcal{Z}^n \to \mathbb{R}$, 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.
- The population cost function: $\overline{\mathcal{L}}(\theta) := \mathbb{E}[\mathcal{L}_n(\theta; Z_1^n)]$. Implicit in this definition is that $\mathcal{L}_n(\theta; Z_1^n) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(\theta; Z_i)$, where $\mathcal{L} : \Omega \times \mathcal{Z} \to \mathbb{R}$ denotes the cost function of one sample.
- The target parameter: $\theta^* = \arg\min_{\theta \in \Omega} \overline{\mathcal{L}}(\theta)$.
- The M-estimator:

$$\widehat{\theta} \in \arg\min_{\theta \in \Omega} \{\mathcal{L}_n(\theta; Z_1^n) + \lambda_n \Phi(\theta)\}, \tag{1}$$

with penalty (regularizer) $\Phi : \Omega \to \mathbb{R}$, and $\lambda_n > 0$ is a user-defined regularization weight.

Examples

Example 1: Lasso

Quadratic cost of the least square:

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

For the purpose of discovering the sparsity, a good choice of regularizer is $\Phi(\theta) = \sum_{j=1}^{d}|\theta_j|$. and thus

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

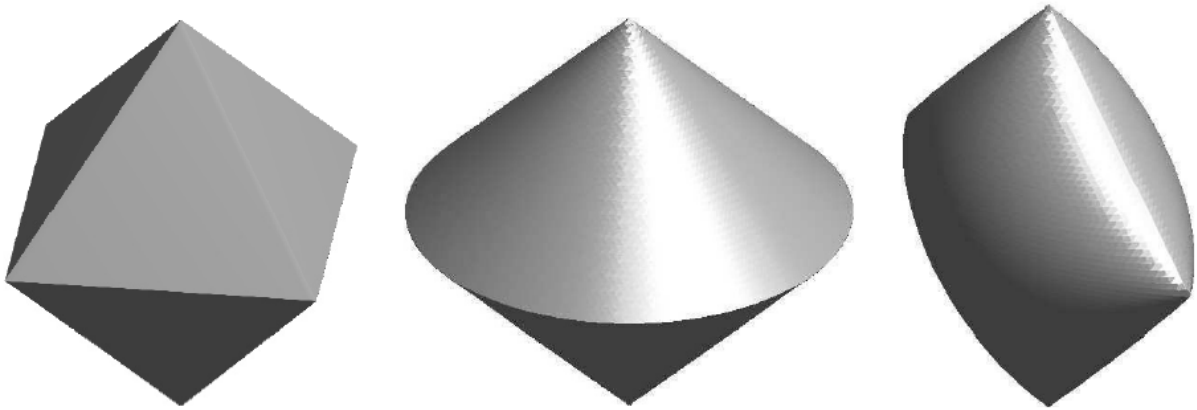
Example

Example 2: Group Lasso

The motivation for group-structured penalties is to estimate parameter vectors whose support lies within a union of a (relatively small) subset of groups. Let $\mathcal{G} = \{g_1, \ldots, g_T\}$ be a disjoint partition of the index set, $[d] := \{1, \ldots, d\}$ — that is, each group g_j is a subset of the index set. When the g_j's are disjoint and covers $[d]$, we can construct the following regularizer:

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

Example



(a) (b) (c)

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 $\Phi(\theta) = \sqrt{\theta_1^2 + \theta_2^2} + \sqrt{\theta_1^2 + \theta_3^2}$.

Preliminary and assumption

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. Specifically,

- the space $\mathbb{R}^d$ with the usual Euclidean inner product, or more generally with a weighted Euclidean inner product
- the space $\mathbb{R}^{d_1 \times d_2}$ equipped with the trace inner product $\langle\langle \mathbf{A}, \mathbf{B} \rangle\rangle := \operatorname{trace}\left(\mathbf{A}^{\mathrm{T}}\mathbf{B}\right) = \sum_{j_1=1}^{d_1} \sum_{j_2=1}^{d_2} A_{j_1 j_2} B_{j_1 j_2}$.

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

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

Definitions

Definition 1

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

$$\Phi(\alpha + \beta) = \Phi(\alpha) + \Phi(\beta) \quad \text{for all } \alpha \in \mathbb{M} \text{ and } \beta \in \bar{\mathbb{M}}^{\perp}.$$

Examples

Example 3: Decomposability and group sparse norms

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

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

corresponding to the set of all vectors that are supported on S. Observe that $\mathbb{M}^{\perp}(S) = \left\{ \theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \in S \right\}$. 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 $\left(\mathbb{M}(S), \mathbb{M}^{\perp}(S)\right)$.

Examples

Example 4: Decomposability and group sparse norms

Given any subset $S_{\mathcal{G}} \subset \mathcal{G}$ of the group index set, consider the set

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

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 $\big(\mathbb{M}(S_{\mathcal{G}}), \mathbb{M}^{\perp}(S_{\mathcal{G}})\big)$.

A key consequence of decomposability

We aim to show that decomposability - in conjunction with a suitable choice for the regularization weight λ_n– ensures that the error $\widehat{\Delta} := \widehat{\theta} - \theta^*$ must lie in a very restricted set.

Definition 2: Dual norm

Given any norm $\Phi : \mathbb{R}^d \to \mathbb{R}$, its dual norm is defined in a variational manner as

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

A key consequence of decomposability

Define the score function as the gradient of the empirical cost evaluated at θ^*: $\nabla \mathcal{L}_n(\theta^*)$.

Intuition:

- Under mild regularity conditions, we have $\mathbb{E}[\nabla \mathcal{L}_n(\theta^*))] = \nabla \overline{\mathcal{L}}(\theta^*)$.
- when the target parameter θ^* lies in the interior of the parameter space Ω, by the optimality conditions for the minimization, the random vector $\nabla \mathcal{L}_n(\theta^*)$ has zero mean.

Thus, 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"

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

A key consequence of decomposability

Proposition 1

Let $\mathcal{L}_n : \Omega \to \mathbb{R}$ be a convex function, let the regularizer $\Phi : \Omega \to [0, \infty)$ be a norm, and consider a subspace pair $\left(\mathbb{M}, \bar{\mathbb{M}}^{\perp}\right)$ 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

$$\mathbb{C}_{\theta^*}\left(\mathbb{M}, \bar{\mathbb{M}}^{\perp}\right) := \left\{\Delta \in \Omega \mid \Phi\left(\Delta_{\bar{\mathbb{M}}^{\perp}}\right) \leq 3\Phi\left(\Delta_{\bar{\mathbb{M}}}\right) + 4\Phi\left(\theta^*_{\mathbb{M}^{\perp}}\right)\right\} \tag{2}$$

An illustration

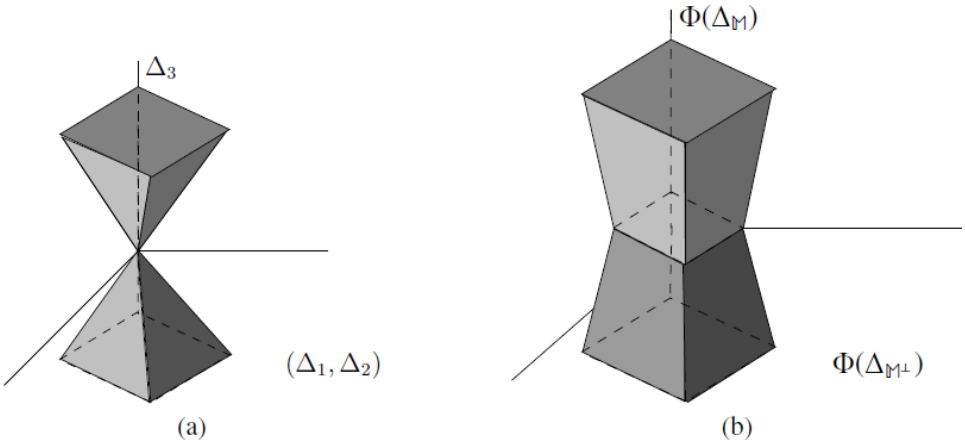


Figure 9.7 Illustration of the set $\mathbb{C}_{\theta^*}(\mathbb{M}, \bar{\mathbb{M}}^\perp)$ in the special case $\Delta = (\Delta_1, \Delta_2, \Delta_3) \in \mathbb{R}^3$ and regularizer $\Phi(\Delta) = \|\Delta\|_1$, relevant for sparse vectors (Example 9.1). This picture shows the case $S = \{3\}$, 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.

A key consequence of decomposability

For showing proposition 1, we need the following lemma:

Lemma 1: Deviation inequalities

For any decomposable regularizer and parameters θ^* and Δ, we have

$$\Phi(\theta^* + \Delta) - \Phi(\theta^*) \geq \Phi\left(\Delta_{\bar{\mathbb{M}}^{\perp}}\right) - \Phi\left(\Delta_{\bar{\mathbb{M}}}\right) - 2\Phi\left(\theta^*_{\mathbb{M}^{\perp}}\right) \tag{3}$$

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

$$\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) \geq -\frac{\lambda_n}{2}\left[\Phi\left(\Delta_{\bar{\mathbb{M}}}\right) + \Phi\left(\Delta_{\bar{\mathbb{M}}^{\perp}}\right)\right] \tag{4}$$

Proof of Lemma 1

Proof.

By triangle ineq:

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

By decomposability,

$$\Phi\left(\theta^*_{\mathbb{M}} + \Delta_{\bar{\mathbb{M}}^\perp}\right) = \Phi\left(\theta^*_{\mathbb{M}}\right) + \Phi\left(\Delta_{\bar{\mathbb{M}}^\perp}\right). \tag{5}$$

Thus, $\Phi\left(\theta^* + \Delta\right) \geq \Phi\left(\theta^*_{\mathbb{M}}\right) + \Phi\left(\Delta_{\bar{\mathbb{M}}^\perp}\right) - \Phi\left(\theta^*_{\mathbb{M}^\perp}\right) - \Phi\left(\Delta_{\bar{\mathbb{M}}}\right)$.
Again by tri-ineq, $\Phi\left(\theta^*\right) \leq \Phi\left(\theta^*_{\mathbb{M}}\right) + \Phi\left(\theta^*_{\mathbb{M}^\perp}\right)$.

Proof of Lemma 1

Proof.

Combine this with (5):

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

which yields (3). By convexity of cost function $\mathcal{L}_n$, we have

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

Applying the Hölder inequality with the regularizer and its dual

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

Proof of Lemma 1

Proof.

along with the assumed bound $\lambda_n \geq 2\Phi^*\left(\nabla \mathcal{L}_n\left(\theta^*\right)\right)$. Putting together the pieces yields the claimed bound (4). □

With the result of Lemma 1, we can proceed to show Proposition 1.

Proof of Proposition 1

Proof.

Consider $\mathcal{F} : \Omega \to \mathbb{R}$ given by

$$\mathcal{F}(\Delta) := \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) + \lambda_n \{\Phi(\theta^* + \Delta) - \Phi(\theta^*)\} \quad (6)$$

the optimality of $\widehat{\theta}$ implies $\widehat{\Delta} = \widehat{\theta} - \theta^*$ must satisfy the condition $\mathcal{F}(\widehat{\Delta}) \leq 0$. Combining (3) and (4):

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

Then the claim follows.

Curvature estimation challenge in High-Dim

The classical role of curvature in MLE: under i.i.d. sampling, MLE is to minimize $\mathcal{L}_n(\theta) := -\frac{1}{n}\sum_{i=1}^{n} \log \mathbb{P}_\theta(z_i)$. The Hessian of this cost function $\nabla^2 \mathcal{L}_n(\theta)$ is the sample version of the Fisher information. With d fixed and sample size $n \to \infty$, the Hessian would converge to the population Fisher information $\nabla^2 \overline{\mathcal{L}}(\theta)$.

The role:

- provides a lower bound on the accuracy of any statistical estimator via the CR 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^*.

Curvature estimation challenge in High-Dim

In the high-dim setting where $n < d$, the sample Fisher information matrix $\nabla^2 \mathcal{L}_n(\theta^*)$ is rank-degenerate. 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*. Two ways to develop such notion: (i) lower bounding the error in the first-order Taylor-series expansion; (ii) lower bounding the curvature of the gradient mapping.

Curvature estimation challenge in High-Dim

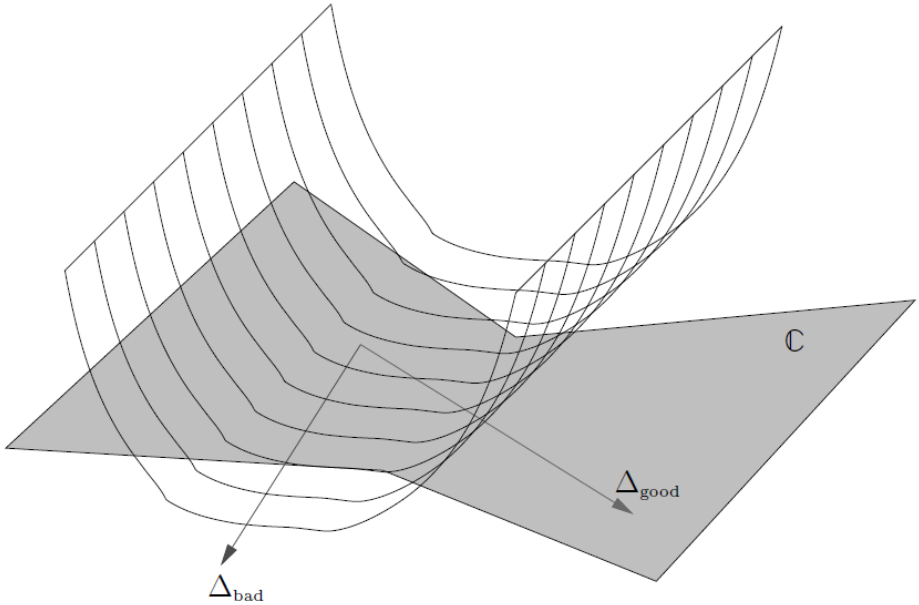


Figure 9.8 Illustration of the cost function $\theta \mapsto \mathcal{L}_n(\theta; Z_1^n)$. In the high-dimensional setting ($d > n$), although it may be curved in certain directions (e.g., Δ_{good}), there are $d - n$ directions in which it is flat up to second order (e.g., Δ_{bad}).

Restricted strong convexity

Firstly, for the 1^{st} order Taylor series error

$$\mathcal{E}_n(\Delta) := \mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle \tag{7}$$

If $\theta \mapsto \mathcal{L}_n(\theta)$ is convex, this error is positive.

Definition 3: κ-strongly convex

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

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2 \tag{8}$$

for all Δ in a neighborhood of the origin.

Restricted strong convexity

As shown in previous discussion, such notion of convexity cannot hold in generic high-dim case. However, for decomposable regularizers, as is stated in Proposition 1, the error vector should belong to a special set. This motivates us to define the restricted strong convexity:

Definition 4: Restricted strong convexity

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

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2\Phi^2(\Delta) \quad \text{for all } \Delta \in \mathbb{B}(R) \tag{9}$$

Restricted strong convexity

Example 5: Restricted eigenvalues for least-squares cost

The restricted eigenvalue (RE) conditions can be shown to correspond to a special case of RSC. 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

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

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.

Restricted strong convexity

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 5: 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

$$\Psi(\mathbb{S}) := \sup_{u \in \mathbb{S} \setminus \{0\}} \frac{\Phi(u)}{\|u\|}. \tag{10}$$

Restricted strong convexity

The Subspace Lipschitz constant 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 $\Delta \in \mathbb{C}_{\theta^*}\left(\mathbb{M}, \bar{\mathbb{M}}^{\perp}\right)$, we have

$$\Phi(\Delta) \overset{\text{(i)}}{\leq} \Phi\left(\Delta_{\bar{\mathbb{M}}}\right) + \Phi\left(\Delta_{\bar{\mathbb{M}}^{\perp}}\right) \overset{\text{(ii)}}{\leq} 4\Phi\left(\Delta_{\bar{\mathbb{M}}}\right) \overset{\text{(iii)}}{\leq} 4\Psi(\bar{\mathbb{M}})\|\Delta\| \tag{11}$$

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

Guarantees under restricted strong convexity

Assumptions through out this section:

- (A1) The cost function is convex, and satisfies the local RSC condition 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 $\mathbb{M} \subseteq \bar{\mathbb{M}}$ such that the regularizer decomposes over $\left(\mathbb{M}, \bar{\mathbb{M}}^{\perp}\right)$.

The "good" event: $\mathbb{G}(\lambda_n) := \left\{\Phi^*\left(\nabla \mathcal{L}_n\left(\theta^*\right)\right) \leq \frac{\lambda_n}{2}\right\}$
Define the quantity:

$$\varepsilon_n^2\left(\bar{\mathbb{M}}, \mathbb{M}^{\perp}\right) := \underbrace{9\frac{\lambda_n^2}{\kappa^2}\Psi^2(\bar{\mathbb{M}})}_{\text{estimation error}} + \underbrace{\frac{8}{\kappa}\left\{\lambda_n \Phi\left(\theta^*_{\mathbb{M}^{\perp}}\right) + 16\tau_n^2 \Phi^2\left(\theta^*_{\mathbb{M}^{\perp}}\right)\right\}}_{\text{approximation error}}, \tag{12}$$

Guarantees under restricted strong convexity

Theorem 1: Bounds for general models

Under conditions (A1) and (A2), consider the regularized M-estimator (1) conditioned on the event $\mathbb{G}(\lambda_n)$
(a) Any optimal solution satisfies the bound

$$\Phi\left(\widehat{\theta}-\theta^*\right) \leq 4\left\{\Psi(\bar{\mathbb{M}})\left\|\widehat{\theta}-\theta^*\right\|+\Phi\left(\theta^*_{\mathbb{M}^\perp}\right)\right\} \tag{13}$$

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

$$\left\|\widehat{\theta}-\theta^*\right\|^2 \leq \varepsilon_n^2\left(\bar{\mathbb{M}}, \mathbb{M}^\perp\right). \tag{14}$$

Guarantees under restricted strong convexity

Remark on Theorem 1

Probabilistic guarantees:

- the RSC condition holds with high probability
- for a concrete choice of regularization parameter, the dual norm bound $\lambda_n \geq 2\Phi^*\left(\nabla \mathcal{L}_n\left(\theta^*\right)\right)$ defining the event $\mathbb{G}\left(\lambda_n\right)$ holds with high probability.

Trade off between two errors: 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.

Guarantees under restricted strong convexity

In the special case that the target parameter θ^* is contained within a subspace $\mathbb{M}$, Theorem 1 has the following corollary:

Corollary 1

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 (1) satisfies the bounds

$$\Phi\left(\widehat{\theta} - \theta^*\right) \leq 6\frac{\lambda_n}{\kappa}\Psi^2(\bar{\mathbb{M}}) \tag{15}$$

$$\left\|\widehat{\theta} - \theta^*\right\|^2 \leq 9\frac{\lambda_n^2}{\kappa^2}\Psi^2(\bar{\mathbb{M}}) \tag{16}$$

Proof of Theorem 1

Proof.

We begin by proving part (a). Letting $\widehat{\Delta} = \widehat{\theta} - \theta^*$ be the error. By tri-ineq

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

where inequality (i) follows from Proposition 1 under event $\mathbb{G}(\lambda_n)$ and inequality (ii) follows from the definition of the optimal subspace constant.

Proof of Theorem 1

Proof.

Adopt the shorthand $\mathbb{C}$ for the set $\mathbb{C}_{\theta^*}\left(\mathbb{M}, \bar{\mathbb{M}}^{\perp}\right)$. 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 (6) over the set $\mathbb{K}(\delta) := \mathbb{C} \cap \{\|\Delta\| = \delta\}$.

Lemma 2

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

Proof of Lemma 2

Proof.

We can prove the contrapositive statement. Assume $\|\widehat{\Delta}\| = \|\widehat{\theta} - \theta^*\| > \delta$, since $\mathbb{C}$ is star-shaped around the origin, thus $t^*\widehat{\Delta} = \left(\delta/\|\widehat{\Delta}\|\right)\widehat{\Delta}$. By convexity of $\mathcal{F}$

$$\begin{aligned}\mathcal{F}\left(t^*\widehat{\Delta}\right) = \mathcal{F}\left(t^*\widehat{\Delta} + (1-t^*)\,0\right) &\leq t^*\mathcal{F}(\widehat{\Delta}) + (1-t^*)\,\mathcal{F}(0) \\ &\overset{\text{(i)}}{=} t^*\mathcal{F}(\widehat{\Delta})\end{aligned}$$

where equality (i) uses the fact that $\mathcal{F}(0) = 0$ by construction. But since $\widehat{\Delta}$ is optimal, we must have $\mathcal{F}(\widehat{\Delta}) \leq 0$, and hence $\mathcal{F}(t^*\Delta) \leq 0$ as well. □

Proof of Theorem 1

Proof.

Fix some radius $\delta \in (0, R]$, on the basis of Lemma 2, it suffice to establish a lower bound on $\mathcal{F}(\Delta)$ for $\forall \Delta \in \mathbb{K}(\delta)$.

$$\begin{aligned}
\mathcal{F}(\Delta) =& \mathcal{L}_n\left(\theta^* + \Delta\right) - \mathcal{L}_n\left(\theta^*\right) + \lambda_n\left\{\Phi\left(\theta^* + \Delta\right) - \Phi\left(\theta^*\right)\right\} \\
\overset{\text{(RSC)}}{\geq}& \left\langle \nabla \mathcal{L}_n\left(\theta^*\right), \Delta \right\rangle + \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2 \Phi^2(\Delta) \\
& + \lambda_n\left\{\Phi\left(\theta^* + \Delta\right) - \Phi\left(\theta^*\right)\right\} \\
\overset{\text{(Bound (3))}}{\geq}& \left\langle \nabla \mathcal{L}_n\left(\theta^*\right), \Delta \right\rangle + \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2 \Phi^2(\Delta) \\
& + \lambda_n\left\{\Phi\left(\Delta_{\bar{\mathbb{M}}^\perp}\right) - \Phi\left(\Delta_{\bar{\mathbb{M}}}\right) - 2\Phi\left(\theta^*_{\mathbb{M}^\perp}\right)\right\}
\end{aligned} \tag{17}$$

Proof of Theorem 1

Proof.

By $|\langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle| \overset{\text{(Hölder)}}{\leq} \Phi^*(\nabla \mathcal{L}_n(\theta^*)) \Phi(\Delta) \overset{(\mathbb{G}(\lambda_n))}{\leq} \frac{\lambda_n}{2} \Phi(\Delta)$,

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

By tri-ineq: $\Phi(\Delta) = \Phi(\Delta_{\bar{\mathbb{M}}^\perp} + \Delta_{\bar{\mathbb{M}}}) \leq \Phi(\Delta_{\bar{\mathbb{M}}^\perp}) + \Phi(\Delta_{\bar{\mathbb{M}}})$,

$$\mathcal{F}(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2 \Phi^2(\Delta) - \frac{\lambda_n}{2}\left\{ 3\Phi(\Delta_{\bar{\mathbb{M}}}) + 4\Phi\left(\theta^*_{\mathbb{M}^\perp}\right) \right\} \quad (18)$$

Proof of Theorem 1

Proof.

Similarly, $\forall \Delta \in \mathbb{C}$

$$\begin{aligned}\Phi^2(\Delta) &\leq \left\{4\Phi\left(\Delta_{\bar{\mathbb{M}}}\right) + 4\Phi\left(\theta^*_{\mathbb{M}^\perp}\right)\right\}^2 \leq 32\Phi^2\left(\Delta_{\bar{\mathbb{M}}}\right) + 32\Phi^2\left(\theta^*_{\mathbb{M}^\perp}\right) \\ &\leq 32\Psi^2(\bar{\mathbb{M}})\|\Delta\|^2 + 32\Phi^2\left(\theta^*_{\mathbb{M}^\perp}\right) \text{ (Subspace Lip+Proj)}\end{aligned} \tag{19}$$

Combine this with (18),

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

Proof of Theorem 1

Proof.

where step (ii) uses the assumed bound $\tau_n^2 \Psi^2(\bar{\mathbb{M}}) < \frac{\kappa}{64}$.
RHS of above ineq is strictly positive definite quadratic form in $\|\Delta\|$, if

$$\|\Delta\|^2 \geq \varepsilon_n^2 \left(\bar{\mathbb{M}}, \mathbb{M}^\perp\right) := 9\frac{\lambda_n^2}{\kappa^2}\Psi^2(\bar{\mathbb{M}}) + \frac{8}{\kappa}\left\{\lambda_n \Phi\left(\theta^*_{\mathbb{M}^\perp}\right) + 16\tau_n^2 \Phi^2\left(\theta^*_{\mathbb{M}^\perp}\right)\right\} \tag{20}$$

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

Bounds under Φ^*-curvature

- An alternative way of characterizing strong convexity of a differentiable cost function is via the behavior of its gradient.
- $\mathcal{L}_n$ is locally κ-strongly convex at θ^*, iff.

$$\langle \nabla \mathcal{L}_n (\theta^* + \Delta)) - \nabla \mathcal{L}_n (\theta^*), \Delta \rangle \geq \kappa \|\Delta\|^2 \tag{21}$$

for all Δ in some ball around zero.

Definition 6

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

$$\Phi^* (\nabla \mathcal{L}_n (\theta^* + \Delta) - \nabla \mathcal{L}_n (\theta^*)) \geq \kappa \Phi^*(\Delta) - \tau_n \Phi(\Delta) \tag{22}$$

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

Bounds under Φ^*-curvature

Example 6: Restricted curvature for least-squares cost

For least square cost function, $\nabla \mathcal{L}_n(\theta) = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}\,(\theta - \theta^*) = \widehat{\mathbf{\Sigma}}\,(\theta - \theta^*)$, where $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X}$. For the ℓ_1-norm as the regularizer Φ, the dual norm Φ^* is the ℓ_∞-norm. Thus the lower bound (22) here is

$$\|\widehat{\Sigma}\Delta\|_\infty \geq \kappa\|\Delta\|_\infty - \tau_n\|\Delta\|_1 \quad \text{for all } \Delta \in \mathbb{R}^d \tag{23}$$

This is closely related to ℓ_∞-restricted eigenvalues of the sample covariance matrix $\widehat{\mathbf{\Sigma}}$:

$$\|\widehat{\Sigma}\Delta\|_\infty \geq \kappa'\|\Delta\|_\infty \quad \text{for all } \Delta \in \mathbb{C}(S;\alpha) \tag{24}$$

where $\mathbb{C}(S;\alpha) := \left\{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \alpha\,\|\Delta_S\|_1\right\}$, and (κ', α) are given positive constants. Ex. 9.11 shows certain equivalence.

Bounds under Φ^*-curvature

Under the following assumptions, we can proceed to show the main result

- (A1′) The cost satisfies the Φ^*-curvature condition (22) with parameters $(\kappa, \tau_n; R)$.
- (A2) The regularizer is decomposable with respect to the subspace pair $\left(\mathbb{M}, \bar{\mathbb{M}}^{\perp}\right)$ with $\mathbb{M} \subseteq \bar{\mathbb{M}}$.

Bounds under Φ^*-curvature

Theorem 2

Given a target parameter $\theta^* \in \mathbb{M}$, consider the regularized M-estimator (1) under conditions (A1′) and (A2), and suppose that $\tau_n \Psi^2(\bar{\mathbb{M}}) < \frac{\kappa}{32}$. Conditioned on the event $\mathbb{G}(\lambda_n) \cap \left\{ \Phi^*\left(\widehat{\theta} - \theta^*\right) \leq R \right\}$, any optimal solution $\widehat{\theta}$ satisfies the bound

$$\Phi^*\left(\widehat{\theta} - \theta^*\right) \leq 3\frac{\lambda_n}{\kappa} \tag{25}$$

Proof of Theorem 2

Proof.

- $\forall$ optimum $\widehat{\theta}$, $\exists$ a subgradient vector $\widehat{z} \in \partial\Phi(\widehat{\theta})$ s.t. $\nabla\mathcal{L}_n(\widehat{\theta}) + \lambda_n\widehat{z} = 0$.
- $\nabla\mathcal{L}_n\left(\theta^* + \widehat{\Delta}\right) - \nabla\mathcal{L}_n\left(\theta^*\right) = -\nabla\mathcal{L}_n\left(\theta^*\right) - \lambda_n\widehat{z}$
- Taking the Φ^*-norm of both sides, use tri-ineq:

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

- By Ex. 9.6 $\Phi^*(\widehat{z}) \leq 1$, on $\mathbb{G}\left(\lambda_n\right)$ where $\Phi^*\left(\nabla\mathcal{L}_n\left(\theta^*\right)\right) \leq \frac{\lambda_n}{2}$,$\Rightarrow$ $\Phi^*\left(\nabla\mathcal{L}_n\left(\theta^* + \Delta\right) - \nabla\mathcal{L}_n\left(\theta^*\right)\right) \leq \frac{3\lambda_n}{2}$.
-

$$\text{By (22): } \kappa\Phi^*(\widehat{\Delta}) \leq \frac{3}{2}\lambda_n + \tau_n\Phi(\widehat{\Delta}). \tag{26}$$

Proof of Theorem 2

- It remains to bound $\Phi(\widehat{\Delta})$ in terms of the dual norm $\Phi^*(\widehat{\Delta})$.

Lemma 3

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

$$\Phi(\Delta) \leq 16\Psi^2(\bar{\mathbb{M}})\Phi^*(\Delta) \quad \text{for any } \Delta \in \mathbb{C}_{\theta^*}\left(\mathbb{M}, \bar{\mathbb{M}}^{\perp}\right) \tag{27}$$

- On $\mathbb{G}(\lambda_n)$, Proposition 14 guarantees that $\widehat{\Delta} \in \mathbb{C}_{\theta^*}\left(\mathbb{M}, \bar{M}^{\perp}\right)$
- Apply bound (27) to $\widehat{\Delta}$, then substituting to (26) $\Rightarrow$ $\left(\kappa - 16\Psi^2(\mathbb{M})\tau_n\right)\Phi^*(\widehat{\Delta}) \leq \frac{3}{2}\lambda_n$
- The claim then follows with the assumption $\Psi^2(\mathbb{M})\tau_n \leq \frac{k}{32}$

Proof of Lemma 3

Proof.

- By (11), if $\theta^* \in \mathbb{M}$, $\Delta \in \mathbb{C}_{\theta^*}\left(\mathbb{M}, \bar{\mathbb{M}}^{\perp}\right) \Rightarrow \Phi(\Delta) \leq 4\Psi(\bar{\mathbb{M}})\|\Delta\|$
- $\|\Delta\|^2 \leq \Phi(\Delta)\Phi^*(\Delta) \leq 4\Psi(\bar{\mathbb{M}})\|\Delta\|\Phi^*(\Delta)$ (Hölder's ineq).
- This is exactly $\|\Delta\| \leq 4\Psi(\bar{\mathbb{M}})\Phi^*(\Delta)$.
- Put together we have:

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

Thanks!