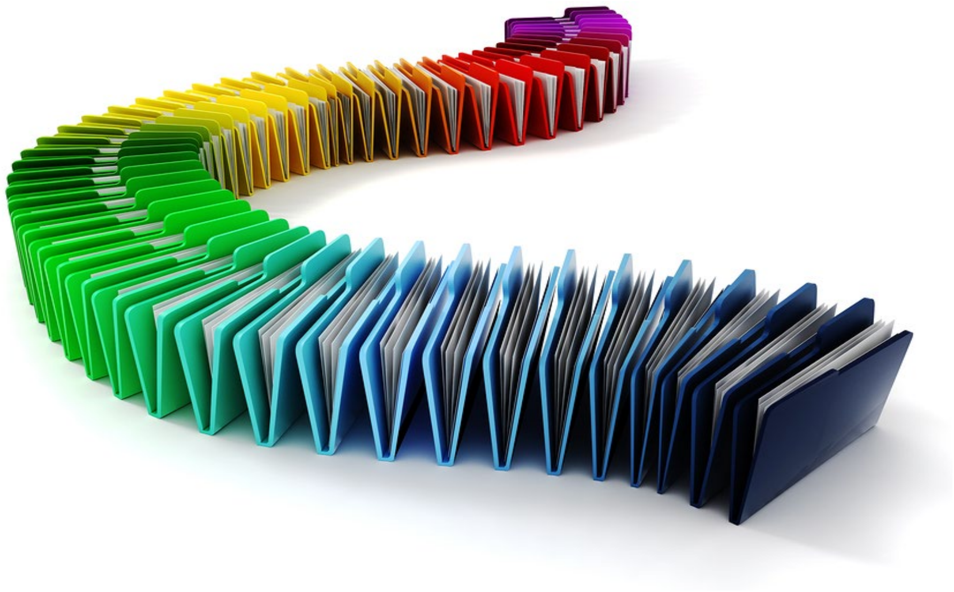
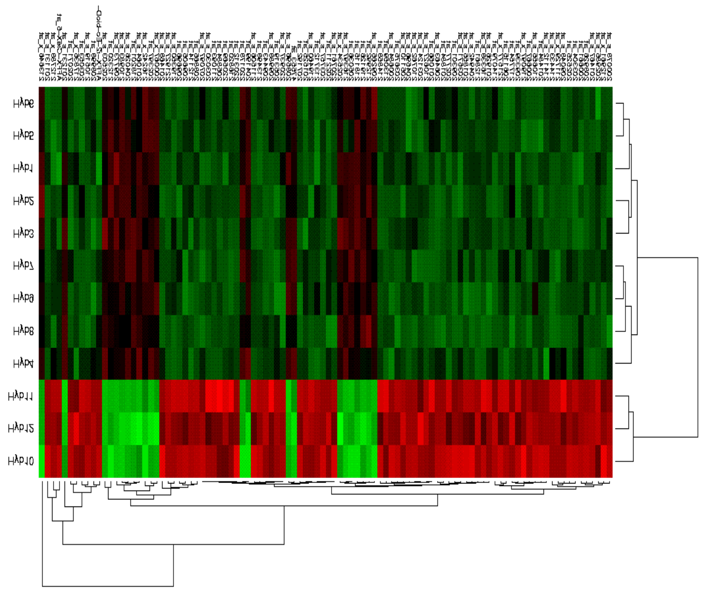
Nonconvex Theory of M-estimators with Decomposable Regularizers

Introduction





- Many real-world problems, like documents and image data have millions of features.
- These data sets appear to have a “high dimensional flavor”, with dimension d larger than the sample size n.
- For many of these applications, classical “large n, fixed d” theory fails to provide useful predictions.

Introduction

- The expectation of loss function $\mathcal{L}_n(\theta; Z_1^n)$ is defined as $\bar{\mathcal{L}}(\theta) := \mathbb{E}(\mathcal{L}_n(\theta; Z_1^n))$. The target parameter θ^* is defined as $\theta^* = \underset{\theta\in\mathbb{R}^d}{\operatorname{argmin}}\, \bar{\mathcal{L}}(\theta)$. The M-estimator is defined as $\hat{\theta} \in \underset{\theta\in\mathbb{R}^d}{\operatorname{argmin}}\, \mathcal{L}_n(\theta; Z_1^n) + \lambda_n \Phi(\theta)$, where $\Phi(\theta)$ is a regularizer or penalty function, λ_n is a user-defined regularization weight, the "M" stands for minimization (or maximization).
- **If dimension d is fixed, sample size n goes to infinity, we have** $\lim_{n\to\infty} \nabla^2 \mathcal{L}_\mathrm{n} = \nabla^2 \bar{\mathcal{L}}$, based on Cramer-Rao Bound, we know the Fisher information matrix $\nabla^2 \bar{\mathcal{L}}$ evaluated at θ^* provides a lower bound on the accuracy of any statistical estimator
- **If $d \geq n$,** $\lim_{n\to\infty} \nabla^2 \mathcal{L}_\mathrm{n} \neq \nabla^2 \bar{\mathcal{L}}$ **, we can not use** Fisher information matrix to get the lower bound.

Decomposability and restricted strong convexity

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

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

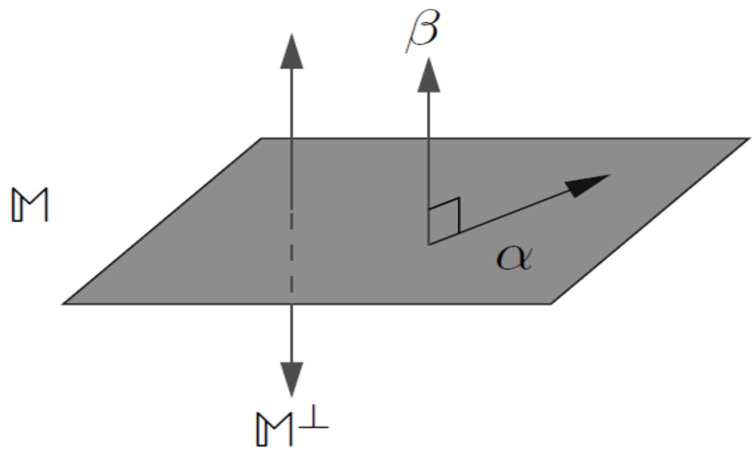


Figure 9.6 In the ideal case, decomposability is defined in terms of a subspace pair $(\mathbb{M}, \mathbb{M}^{\perp})$. For any $\alpha \in \mathbb{M}$ and $\beta \in \mathbb{M}^{\perp}$, the regularizer should decompose as $\Phi(\alpha + \beta) = \Phi(\alpha) + \Phi(\beta)$.

Wainwright, M. High-dimensional statistics: A nonasymptotic viewpoint. Cambridge University Press, 2019.

Decomposability and restricted strong convexity

Proposition 9.13 *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 $(\mathbb{M}, \bar{\mathbb{M}}^{\perp})$ 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*

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

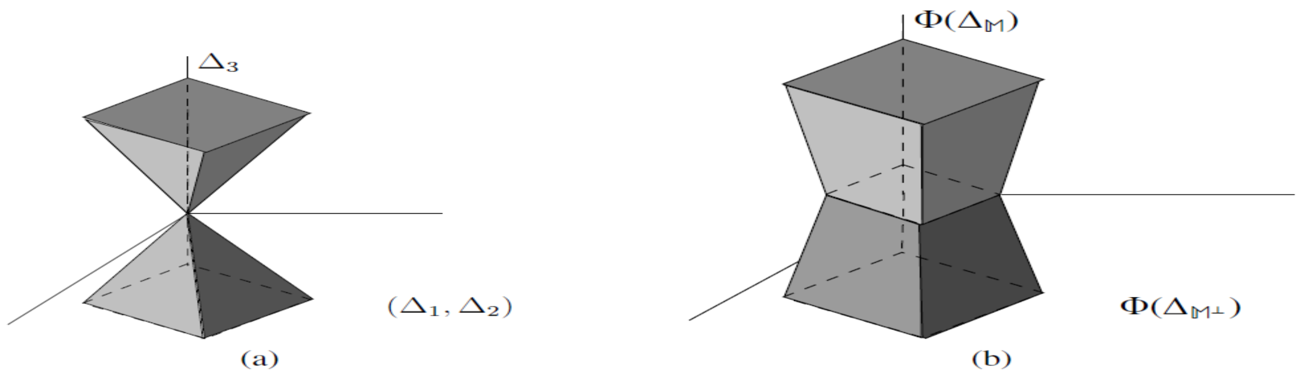


Figure 9.7 Illustration of the set $\mathbb{C}_{\theta^*}(\mathbb{M}, \bar{\mathbb{M}}^{\perp})$ in the special case $\Delta = (\Delta_1, \Delta_2, \Delta_3) \in \mathbb{R}^3$ and regularizer $\Phi(\Delta) = \|\Delta\|_1$, relevant for sparse vectors (Example 9.1). This picture shows the case $S = \{3\}$,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

Wainwright, M. High-dimensional statistics: A nonasymptotic viewpoint. Cambridge University Press, 2019.

Decomposability and 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

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

- 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

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

Wainwright, M. High-dimensional statistics: A nonasymptotic viewpoint. Cambridge University Press, 2019.

Guarantees under restricted strong convexity

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*

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

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

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

(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 $\mathbb{M} \subseteq \bar{\mathbb{M}}$ such that the regularizer decomposes over $(\mathbb{M}, \bar{\mathbb{M}}^\perp)$.

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

Wainwright, M. High-dimensional statistics: A nonasymptotic viewpoint. Cambridge University Press, 2019.

Questions

- (1)Whether the results of Proposition 9.13 in (Wainwright, 2019) still hold if the loss function is nonconvex?

- (2)Can we recover the convergence rates of the estimation error $\|\hat{\theta} - \theta^*\|^2$(9.48b) in (Wainwright, 2019) if the loss function is nonconvex?

Main Contribution

- Stationary points $\hat{\theta} \in \mathbb{R}^d$: $\langle \nabla \mathcal{L}_n(\hat{\theta}) + \lambda_n \nabla \Phi(\hat{\theta}), \theta - \hat{\theta} \rangle \geq 0, \theta \in \mathbb{R}^d$ (1) $\widetilde{\mathbb{G}}(\lambda_n) := \{\Phi^*(\nabla \mathcal{L}_n(\hat{\theta})) \leq \lambda_n/2\}$

- **Theorem1**: Consider any vector $\hat{\theta} \in \mathbb{R}^d$ satisfies (1), conditioned on the event $\widetilde{\mathbb{G}}(\lambda_n)$, we have $\boldsymbol{\hat{\theta} - \theta^*} \in \mathbb{C} := \{\Delta \in \mathbb{R}^{\boldsymbol{d}} | \boldsymbol{\Phi(\Delta_{\bar{M}^\perp}) \leq 3\Phi(\Delta_{\bar{M}}) + 4\Phi(\theta^*_{M^\perp})}\}$

- Remark. Theorem1 shows that the results of the Proposition 9.13 in (Wainwright, 2019) **still hold for any stationary points**. But we have to pay the price. The price is to redefine $\widetilde{\mathbb{G}}(\lambda_n)$ on $\hat{\theta}$ instead of θ^*.

Main Contribution

- Weaker RSC condition: $\langle \nabla \mathcal{L}(\theta^* + \Delta) - \nabla \mathcal{L}(\theta^*), \Delta \rangle \geq \kappa ||\Delta||^2 - \tau_n^2 \Phi^2(\Delta)$(2)

- **Theorem2**:Suppose the loss function satisfies (2). Consider any vector $\hat{\theta} \in \mathbb{R}^d$ satisfies (1), conditioned on the event $\widetilde{\mathbb{G}}(\lambda_n)$, if $\tau_n^2 \Psi^2(\bar{M}) \leq \frac{\kappa}{128}$,we have $||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*||^2 \leq \boldsymbol{\varepsilon}_n^2(\bar{\mathbb{M}}, \mathbb{M}^{\perp})$

- Remark. Theorem 2 shows that **we can still recover the convergence rate of the estimation error under nonconvex condition**, The price is to use the weaker RSC condition and redefined $\widetilde{\mathbb{G}}(\lambda_n)$

Conclusions

- This paper extends the theory of M-estimators with decomposable regularizers from convex to nonconvex

- Theorem 1 recovers the results of the Proposition 9.13 in (Wainwright, 2019) for any stationary points.

- Theorem 2 recovers the convergence rates of the error $\|\hat{\theta} - \theta^*\|^2$(9.48b) in (Wainwright, 2019) for any stationary points.

- Moreover, we use two nonconvex examples to illustrate our main results.

Thank you ！