Algorithmic Foundations of Learning

Lecture 13
The Lasso Estimator. Proximal Gradient Methods

Patrick Rebeschini

Department of Statistics
University of Oxford
November 26th 2019
Convex Recovery: Lasso Estimator

- **Problem:**
  
  \[
  W^0 := \arg\min_{w: \|w\|_0 \leq k} \frac{1}{2n} \|xw - Y\|_2^2
  \]
  is not a convex program

- The set \(\{w \in \mathbb{R}^d : \|w\|_0 \leq k\}\) is not convex

- **Idea:** Use \(\|w\|_1 \leq k\) instead, i.e.,

  \[
  W^1 := \arg\min_{w: \|w\|_1 \leq k} \frac{1}{2n} \|xw - Y\|_2^2
  \]

  ?

- This works, but we look at penalized estimators instead

- Equivalent (in theory!) form of regularization: constrained vs. penalized

- For a given \(\lambda > 0\) (to be tuned):

  \[
  W^{p1} := \arg\min_{w \in \mathbb{R}^d} \min \{R(w) + \lambda \|w\|_1 \}
  \]

- **Lasso estimator:** \(R(w) = \frac{1}{2n} \|xw - Y\|_2^2\)
Convex Recovery. Restricted Strong Convexity

Algorithm:

\[
W_p^1 := \arg\min_{w \in \mathbb{R}^d} R(w) + \lambda \|w\|_1
\]

Restricted strong convexity (Assumption 13.1)

- Function \( R \) convex and differentiable
- \( S = \text{supp}(w^*) := \{i \in [d] : w^*_i \neq 0\} \)
- Cone set: \( \mathcal{C} := \{w \in \mathbb{R}^d : \|w_{S^c}\|_1 \leq 3\|w_S\|_1\} \)  (this is NOT convex!)

There exists \( \alpha > 0 \) such that for any vector \( w \in \mathcal{C} \) we have

\[
R(w^* + w) \geq R(w^*) + \langle \nabla R(w^*), w \rangle + \alpha \|w\|_2^2
\]

Analogue of restricted eigenvalues assumption for \( \ell_0 \) recovery:

- If \( R(w) = \frac{1}{2n} \|xw - Y\|_2^2 \) then \( \nabla R(w) = \frac{1}{n} x^\top (xw - Y) \)
- As \( Y = xw^* + \sigma \xi \), then, for any \( w \in \mathcal{C} \),

\[
\frac{1}{2n} \|xw\|_2^2 \geq \alpha \|w\|_2^2
\]
Convex Recovery. Statistical Guarantees

Statistical Guarantees Convex Recovery (Theorem 13.4)

If the restricted strong convexity assumption holds and \( \lambda \geq 2\|\nabla R(w^*)\|_\infty \), then

\[
\|W^{p1} - w^*\|_2 \leq \frac{3}{2} \frac{\lambda \sqrt{\|w^*\|_0}}{\alpha}
\]

If \( R(w) = \frac{1}{2n} \|xw - Y\|_2^2 \) then \( \|\nabla R(w^*)\|_\infty = \frac{\sigma}{n} \|x^\top \xi\|_\infty \)

- If \( \lambda = 2\|\nabla R(w^*)\|_\infty \), then
  \[
  \|W^{p1} - w^*\|_2 \leq 3 \frac{\sigma \sqrt{\|w^*\|_0}}{\alpha} \frac{\|x^\top \xi\|_\infty}{n}
  \]
- If \( k = \|w^*\|_0 \), then
  \[
  \|W^0 - w^*\|_2 \leq \sqrt{2} \frac{\sigma \sqrt{\|w^*\|_0}}{\alpha} \frac{\|x^\top \xi\|_\infty}{n}
  \]

Same statistical rates (modulo constants). Advantages:

- Convex program! (once again a convex relaxation does not hurt...)
- No need to know sparsity (or upper bounds)
  But we need to known noise (or upper bounds)

Same bounds in expectation and in probability
Restricted Strong Convexity: Sufficient Conditions

In general, checking if restricted strong convexity holds is \textbf{NP hard}

### Tractable Sufficient Conditions for RSC (Proposition 13.5)

- For a matrix $M$, let $\|M\| := \max_{i,j} |M_{ij}|$
- Let $R(w) = \frac{1}{2n} \|xw - Y\|_2^2$
- \[ \| \frac{x^\top x}{n} - I \| \leq \frac{1}{32\|w^*\|_0} \] (Incoherence parameter: $\| \frac{x^\top x}{n} - I \|$)

Then, restricted strong convexity holds with $\alpha = \frac{1}{4}$: $\frac{1}{2n} \|xw\|_2^2 \geq \frac{\|w\|_2^2}{4} \forall w \in C$

### Random Ensembles (Proposition 13.6)

Let $X \in \mathbb{R}^{n \times d}$ with i.i.d. Rademacher r.v.’s. If $n \geq 2048\tau\|w^*\|_0^2 \log d$, $\tau \geq 2$,

\[
\mathbb{P}\left( \left\| \frac{X^\top X}{n} - I \right\| < \frac{1}{32\|w^*\|_0} \right) \geq 1 - \frac{2}{d^{\tau-2}}
\]

Note that $n$ is compared against $\log d$, which is what we want for $n \ll d$
Phase Transitions

**Fundamental limitation:** \( n \gtrsim \|w^*\|_0 \log d \)

From the book “Statistical Learning with Sparsity The Lasso and Generalizations” by Hastie, Tibshirani, Wainwright

**Phase transition** (plot of \( \frac{\|w^*\|_0}{n} \) versus \( \frac{n}{d} \); red = DIFFICULT, blue = EASY)
Computing the Lasso? Proximal Gradient Methods

Lasso estimator: \[
\arg\min_{w \in \mathbb{R}^d} \frac{1}{2n} \|xw - Y\|_2^2 + \lambda \|w\|_1
\]

General structure:
\[
\arg\min_{x \in \mathbb{R}^d} h(x) := f(x) + g(x)
\]
where \( f : \mathbb{R}^d \to \mathbb{R} \) is convex and \( \beta \)-smooth, and \( g : \mathbb{R}^d \to \mathbb{R} \)

Smoothness yields natural algorithm:
\[
h(y) \leq g(y) + f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|_2^2
\]
\[
\arg\min_{y \in \mathbb{R}^d} \left\{ g(y) + f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|_2^2 \right\} = \text{Prox}_{g/\beta} \left( x - \frac{1}{\beta} \nabla f(x) \right)
\]

Proximal operator associated to \( \kappa : \mathbb{R}^d \to \mathbb{R} \):
\[
\text{Prox}_\kappa(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ \kappa(y) + \frac{1}{2} \|y - x\|_2^2 \right\}
\]
Proximal Gradient Methods

Proximal Gradient Method

\[ x_{s+1} = \text{Prox}_{\eta_s g}(x_s - \eta_s \nabla f(x_s)) \]

Proximal Gradient Methods (Theorem 13.8)

- Let \( f \) be convex and \( \beta \)-smooth
- Let \( g \) be convex
- Assume \( \|x_1 - x^*\|_2 \leq b \)

Then, the proximal gradient to minimize \( h = f + g \) with \( \eta_s \equiv \eta = 1/\beta \) satisfies

\[
    h(x_t) - h(x^*) \leq \frac{\beta b^2}{2(t-1)}
\]

- \( O(1/t) \) better than \( O(1/\sqrt{t}) \) of subgradient descent for non-smooth func.
- **Reason:** Beyond first order oracle (need global info on \( g \) to have \( \text{Prox}_{\eta_s g} \))
- Can be accelerated to \( O(1/t^2) \)
Proximal Gradient Methods for the Lasso: ISTA

**Compute Prox?** It reduces to \( d \) one-dim. problems if \( \kappa \) is decomposable:

\[
\text{Prox}_\kappa(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ \sum_{i=1}^{d} \kappa_i(y_i) + \frac{1}{2} \sum_{i=1}^{d} (y_i - x_i)^2 \right\} = \begin{pmatrix} \text{Prox}_{\kappa_1}(x_1) \\ \vdots \\ \text{Prox}_{\kappa_d}(x_d) \end{pmatrix}
\]

For the Lasso:

\[
\iota(w; \theta) := \text{Prox}_\theta \cdot |(w) = \arg\min_{y \in \mathbb{R}} \left\{ \theta |y| + \frac{1}{2} (y - w)^2 \right\} = \begin{cases} 
  w - \theta & \text{if } w > \theta \\
  0 & \text{if } -\theta \leq w \leq \theta \\
  w + \theta & \text{if } w < -\theta
\end{cases}
\]

**Iterative Shrinkage-Thresholding Algorithm (ISTA)**

\[
W_{s+1} = \iota \left( W_s - \frac{\eta_s}{n} x^\top (x W_s - Y); \lambda \eta_s \right)
\]

\( R \) is \( \beta \)-smooth, \( \beta = \mu_{\max}\left(\frac{1}{n}x^\top x\right) \), but not strongly convex as \( \mu_{\min}\left(\frac{1}{n}x^\top x\right) = 0 \)

**Proximal Gradient Methods (Theorem 13.8)**

\[
R(W_t) + \lambda \|W_t\|_1 - (R(W^{p1}) + \lambda \|W^{p1}\|_1) \leq \beta \frac{\|W_1 - W^{p1}\|_2^2}{2(t - 1)}
\]