Mathematics of Machine Learning - Summer School

Lecture 10
The Lasso Estimator. Proximal Gradient Methods

July 2, 2021

Patrick Rebeschini
Department of Statistics, University of Oxford

Convex Recovery: Lasso Estimator

- **Problem:** $\boxed{W^0 := \underset{w:\|w\|_0 \le k}{\operatorname{argmin}} \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2}$ is **not** a convex program
- The set $\{w \in \mathbb{R}^d : \|w\|_0 \le k\}$ is not convex
- **Idea:** Use $\|w\|_1 \le k$ instead, i.e.,

$$W^1 := \underset{w:\|w\|_1 \le k}{\operatorname{argmin}} \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2 \qquad ?$$

- 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):

$$\boxed{W^{p1} := \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} R(w) + \lambda\|w\|_1}$$

- **Lasso estimator:** $R(w) = \frac{1}{2n}\|\mathbf{x}w - Y\|_2^2$

Convex Recovery. Restricted Strong Convexity

Algorithm: $\boxed{W^{p1} := \underset{w \in \mathbb{R}^d}{\operatorname{argmin}}\, R(w) + \lambda \|w\|_1}$

Restricted strong convexity (Assumption 13.1)

- Function R convex and differentiable
- $S = \operatorname{supp}(w^\star) := \{i \in [d] : w_i^\star \neq 0\}$
- **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

$$\boxed{R(w^\star + w) \geq R(w^\star) + \langle \nabla R(w^\star), w \rangle + \alpha \|w\|_2^2}$$

Analogue of restricted eigenvalues assumption for ℓ_0 recovery:

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

Convex Recovery. Statistical Guarantees

Statistical Guarantees Convex Recovery (Theorem 13.4)

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

$$\boxed{\|W^{p1} - w^\star\|_2 \leq \frac{3}{2}\frac{\lambda\sqrt{\|w^\star\|_0}}{\alpha}}$$

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

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

Same statistical rates (modulo constants). **Advantages:**

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

Same bounds in expectation and in probability

Proof of Theorem 13.4 (Part I)

Let $\Delta = W^{p1} - w^\star$.

- **Part 1: Prove that $\Delta \in \mathcal{C}$.** By convexity of R we have

$$\begin{aligned} 0 &\le R(W^{p1}) - R(w^\star) - \langle \nabla R(w^\star), \Delta \rangle \\ &= R(W^{p1}) + \lambda \|W^{p1}\|_1 - \lambda \|W^{p1}\|_1 - R(w^\star) - \langle \nabla R(w^\star), \Delta \rangle \\ &\le \lambda \|w^\star\|_1 - \lambda \|w^\star + \Delta\|_1 - \langle \nabla R(w^\star), \Delta \rangle, \end{aligned} \tag{1}$$

  where, by the definition of W^{p1}, $R(W^{p1}) + \lambda \|W^{p1}\|_1 \le R(w^\star) + \lambda \|w^\star\|_1$.

- By Hölder's inequality and the fact that the ℓ_1 norm decomposes so that $\|w^\star + \Delta\|_1 = \|w^\star_S + \Delta_S\|_1 + \|w^\star_{S^c} + \Delta_{S^c}\|_1$, and $w^\star_{S^c} = 0$, we get

$$0 \le \lambda \|w^\star\|_1 - \lambda \|w^\star_S + \Delta_S\|_1 - \lambda \|\Delta_{S^c}\|_1 + \|\nabla R(w^\star)\|_\infty \|\Delta\|_1$$

- Using the assumption $\|\nabla R(w^\star)\|_\infty \le \frac{\lambda}{2}$ and the fact that the reverse triangle inequality yields $\|w^\star_S\|_1 - \|\Delta_S\|_1 \le \|w^\star_S + \Delta_S\|_1$, we get

$$0 \le \lambda \|\Delta_S\|_1 - \lambda \|\Delta_{S^c}\|_1 + \frac{\lambda}{2} \|\Delta\|_1 = \frac{3\lambda}{2} \|\Delta_S\|_1 - \frac{\lambda}{2} \|\Delta_{S^c}\|_1. \tag{2}$$

  Rearranging this expression we obtain $3\|\Delta_S\|_1 \ge \|\Delta_{S^c}\|_1$, so $\Delta \in \mathcal{C}$.

Proof of Theorem 13.4 (Part I)

- **Part 2: Prove the inequality.** As $\Delta \in \mathcal{C}$, we can apply the restricted strong convexity assumption, Assumption 13.1, with $w = \Delta$ and we get

$$\alpha \|\Delta\|_2^2 \leq R(W^{p1}) - R(w^\star) - \langle \nabla R(w^\star), \Delta \rangle,$$

which is analogous to (1) with 0 replaced by $\alpha \|\Delta\|_2^2$.

- Following the exact same steps as in Part 1, (2) now becomes

$$\alpha \|\Delta\|_2^2 \leq \frac{3\lambda}{2} \|\Delta_S\|_1 - \frac{\lambda}{2} \|\Delta_{S^c}\|_1.$$

- This yields, by the Cauchy-Schwarz's inequality,

$$\begin{aligned} \alpha \|\Delta\|_2^2 &\leq \frac{3\lambda}{2} \|\Delta_S\|_1 = \frac{3\lambda}{2} \langle \mathrm{sign}(\Delta_S), \Delta_S \rangle \leq \frac{3\lambda}{2} \sqrt{\|w^\star\|_0} \|\Delta_S\|_2 \\ &\leq \frac{3\lambda}{2} \sqrt{\|w^\star\|_0} \|\Delta\|_2, \end{aligned}$$

where we used that the cardinality of S is equal to $\|w^\star\|_0$, and that the ℓ_2 norm of a vector can only increase if we add non-zero coordinates.

Restricted Strong Convexity: Sufficient Conditions

In general, checking if restricted strong convexity holds is **NP hard**

Tractable Sufficient Conditions for RSC (Proposition 13.5)

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

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

Random Ensembles (Proposition 13.6)

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

$$\boxed{\mathbf{P}\left(\left\|\frac{\mathbf{X}^\top \mathbf{X}}{n} - I\right\| < \frac{1}{32\|w^\star\|_0}\right) \geq 1 - \frac{2}{d^{\tau-2}}}$$

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

Proof of Proposition 13.5

- Let $w \in \mathcal{C}$. We have
$$\frac{1}{2n}\|\mathbf{x}w\|_2^2 = \frac{1}{2n}w^\top \mathbf{x}^\top \mathbf{x}w = \frac{1}{2}w^\top(\mathbf{c}-I)w + \frac{\|w\|_2^2}{2}.$$

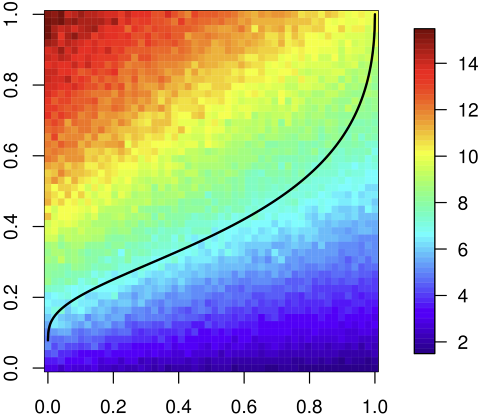
- Recall that Hölder's inequality gives $|a^\top b| \leq \|a\|_1\|b\|_\infty$, or equivalently, $-\|a\|_1\|b\|_\infty \leq a^\top b \leq \|a\|_1\|b\|_\infty$. Applying the lower bound we get,
$$\frac{1}{2n}\|\mathbf{x}w\|_2^2 \geq \frac{\|w\|_2^2}{2} - \frac{\|w\|_1}{2}\|(\mathbf{c}-I)w\|_\infty \geq \frac{\|w\|_2^2}{2} - \frac{\|w\|_1^2}{2}\|\mathbf{c}-I\|.$$

- As $w \in \mathcal{C}$, $\|w_{S^c}\|_1 \leq 3\|w_S\|_1$ and $S = \mathrm{supp}(w^\star) := \{i \in [d] : w_i^\star \neq 0\}$, by the Cauchy-Schwarz's inequality we have
$$\begin{aligned}\|w\|_1 = \|w_S\|_1 + \|w_{S^c}\|_1 \leq 4\|w_S\|_1 = 4\langle \mathrm{sign}(w_S), w_S\rangle \\ \leq 4\sqrt{\|w^\star\|_0}\|w_S\|_2 \leq 4\sqrt{\|w^\star\|_0}\|w\|_2\end{aligned}$$

- Hence, using the assumption of the proposition, we get
$$\frac{1}{2n}\|\mathbf{x}w\|_2^2 \geq \frac{\|w\|_2^2}{2} - 8\|w^\star\|_0\|w\|_2^2\|\mathbf{c}-I\| \geq \frac{\|w\|_2^2}{2} - \frac{\|w\|_2^2}{4} = \frac{\|w\|_2^2}{4}.$$

Phase Transitions

Fundamental limitation: $\boxed{n \gtrsim \|w^\star\|_0 \log d}$



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

Phase transition (plot of $\frac{\|w^\star\|_0}{n}$ versus $\frac{n}{d}$; red = DIFFICULT, blue = EASY)

Computing the Lasso? Proximal Gradient Methods

Lasso estimator: $$\operatorname*{argmin}_{w \in \mathbb{R}^d} \frac{1}{2n} \|\mathbf{x}w - Y\|_2^2 + \lambda \|w\|_1$$

- General structure:

$$\boxed{\operatorname*{argmin}_{x \in \mathbb{R}^d} h(x) := f(x) + g(x)}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and β-smooth, and $g : \mathbb{R}^d \to \mathbb{R}$

- **Smoothness yields natural algorithm:**

$$h(y) \leq g(y) + f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|_2^2$$

$$\operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ g(y) + f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|y - x\|_2^2 \right\} = \mathtt{Prox}_{g/\beta}\left(x - \frac{1}{\beta} \nabla f(x) \right)$$

- **Proximal operator** associated to $\kappa : \mathbb{R}^d \to \mathbb{R}$:

$$\boxed{\mathtt{Prox}_\kappa(x) := \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ \kappa(y) + \frac{1}{2} \|y - x\|_2^2 \right\}}$$

Proximal Gradient Methods

Proximal Gradient Method

$$\boxed{x_{s+1} = \mathtt{Prox}_{\eta_s g}(x_s - \eta_s \nabla f(x_s))}$$

Proximal Gradient Methods (Theorem 13.8)

- Let f be convex and β-smooth
- Let g be convex
- Assume $\|x_1 - x^\star\|_2 \le b$

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

$$\boxed{h(x_t) - h(x^\star) \le \frac{\beta b^2}{2(t-1)}}$$

- $O(1/t)$ 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 $\mathtt{Prox}_{\eta_s g}$)
- Can be accelerated to $O(1/t^2)$

Proximal Gradient Methods for the Lasso: ISTA

Compute $\texttt{Prox}$? It reduces to d one-dim. problems if κ is decomposable:

$$\texttt{Prox}_\kappa(x) := \operatorname*{argmin}_{y\in\mathbb{R}^d}\left\{\sum_{i=1}^d \kappa_i(y_i) + \frac{1}{2}\sum_{i=1}^d (y_i - x_i)^2\right\} = \begin{pmatrix}\texttt{Prox}_{\kappa_1}(x_1)\\ \vdots \\ \texttt{Prox}_{\kappa_d}(x_d)\end{pmatrix}$$

For the Lasso:

$$\iota(w;\theta) := \texttt{Prox}_{\theta|\,\cdot\,|}(w) = \operatorname*{argmin}_{y\in\mathbb{R}}\left\{\theta|y| + \frac{1}{2}(y-w)^2\right\} = \begin{cases} w-\theta & \text{if } w > \theta \\ 0 & \text{if } -\theta \le w \le \theta \\ w+\theta & \text{if } w < -\theta \end{cases}$$

Iterative Shrinkage-Thresholding Algorithm (ISTA)

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

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

Proximal Gradient Methods (Theorem 13.8)

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