# $\ell_1$-Regularized Linear Regression: Persistence and Oracle Inequalities

**Peter Bartlett**
EECS and Statistics
UC Berkeley

slides at http://www.stat.berkeley.edu/∼bartlett

Joint work with Shahar Mendelson and Joe Neeman.

## $\ell_1$-regularized linear regression

- Random pair: $(X, Y) \sim P$, in $\mathbb{R}^d \times \mathbb{R}$.
- $n$ independent samples drawn from $P$: $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- Find $\beta$ so linear function $\langle X, \beta \rangle$ has small risk,

$$P\ell_\beta = P \left( \langle X, \beta \rangle - Y \right)^2 .$$

Here, $\ell_\beta(X, Y) = (\langle X, \beta \rangle - Y)^2$ is the quadratic loss of the linear prediction.

## $\ell_1$-regularized linear regression

- Random pair: $(X, Y) \sim P$, in $\mathbb{R}^d \times \mathbb{R}$.
- $n$ independent samples drawn from $P$: $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- Find $\beta$ so linear function $\langle X, \beta \rangle$ has small risk,

$$P\ell_\beta = P\left(\langle X, \beta \rangle - Y\right)^2 .$$

**Example.** $\ell_1$-regularized least squares:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \left( P_n \ell_\beta + \rho_n \|\beta\|_{\ell_1^d} \right),$$

$$\text{where } P_n \ell_\beta = \frac{1}{n} \sum_{i=1}^{n} \left(\langle X_i, \beta \rangle - Y_i\right)^2, \text{ and } \|\beta\|_{\ell_1^d} = \sum_{j=1}^{d} |\beta_j|.$$

# $\ell_1$-regularized linear regression

**Example.** $\ell_1$-regularized least squares:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \left( P_n \ell_\beta + \rho_n \|\beta\|_{\ell_1^d} \right),$$

$$\text{where } P_n \ell_\beta = \frac{1}{n} \sum_{i=1}^{n} (\langle X_i, \beta \rangle - Y_i)^2 \text{, and } \|\beta\|_{\ell_1^d} = \sum_{j=1}^{d} |\beta_j|.$$

- Tends to select sparse solutions (few non-zero components $\beta_j$).
- Useful, for example, if $d \gg n$.

## $\ell_1$-regularized linear regression

**Example.** $\ell_1$-regularized least squares:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \left( P_n \ell_\beta + \rho_n \|\beta\|_{\ell_1^d} \right),$$

**Example.** $\ell_1$-constrained least squares:

$$\hat{\beta} = \arg\min_{\|\beta\|_{\ell_1^d} \le b_n} P_n \ell_\beta.$$

[Recall: $\ell_\beta(X, Y) = (\langle X, \beta \rangle - Y)^2$.]

# $\ell_1$-regularized linear regression

**Example.** $\ell_1$-regularized least squares:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \left( P_n \ell_\beta + \rho_n \|\beta\|_{\ell_1^d} \right),$$

**Example.** $\ell_1$-constrained least squares:

$$\hat{\beta} = \arg \min_{\|\beta\|_{\ell_1^d} \leq b_n} P_n \ell_\beta.$$

Some questions:

- **Prediction:** Does $\hat{\beta}$ give accurate forecasts?
  e.g., How does $P\ell_{\hat{\beta}}$ compare with $P\ell_{\beta^*}$?

$$\text{Here, } \beta^* = \arg\min \left\{ P\ell_\beta : \|\beta\|_{\ell_1^d} \leq b_n \right\}.$$

# $\ell_1$-regularized linear regression

**Example.** $\ell_1$-regularized least squares:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^d} \left( P_n \ell_\beta + \rho_n \|\beta\|_{\ell_1^d} \right),$$

**Example.** $\ell_1$-constrained least squares:

$$\hat{\beta} = \arg \min_{\|\beta\|_{\ell_1^d} \le b_n} P_n \ell_\beta.$$

Some questions:

- Does $\hat{\beta}$ give accurate forecasts?
  e.g., $P\ell_{\hat{\beta}}$ versus $P\ell_{\beta^*} = \min \left\{ P\ell_\beta : \|\beta\|_{\ell_1^d} \le b_n \right\}$?
- **Estimation:** Under assumptions on $P$, is $\hat{\beta} \approx$ correct?
- **Sparsity Pattern Estimation:** Under assumptions on $P$, are the non-zeros of $\hat{\beta}$ correct?

# Outline of Talk

1. For $\ell_1$-constrained least squares, bounds on $P\ell_{\hat{\beta}} - P\ell_{\beta^*}$.
   - **Persistence:** (Greenshtein and Ritov, 2004)
     For what $d_n, b_n \to \infty$ does $P\ell_{\hat{\beta}} - P\ell_{\beta^*} \to 0$?
   - **Convex Aggregation:** (Tsybakov, 2003)
     For $b = 1$ (convex combinations of dictionary functions), what is rate of $P\ell_{\hat{\beta}} - P\ell_{\beta^*}$?
2. For $\ell_1$-regularized least squares, oracle inequalities.
3. Proof ideas.

## $\ell_1$-regularized linear regression

**Key Issue**: $\ell_\beta$ is unbounded, so some key tools (e.g., concentration inequalities) cannot immediately be applied.

- For $(X, Y)$ bounded, $\ell_\beta$ can be bounded using $\|\beta\|_{\ell_1^d}$, but this gives loose prediction bounds.
- We use chaining to show that metric structures of $\ell_1$-constrained linear functions under $P_n$ and $P$ are similar.

# Main Results: Excess Risk

For $\ell_1$-constrained least squares,

$$\hat{\beta} = \arg\min_{\|\beta\|_{\ell_1^d} \le b} P_n \ell_\beta,$$

if $X$ and $Y$ have suitable tail behaviour then, with probability $1-\delta$,

$$P\ell_{\hat{\beta}} - P\ell_{\beta^*} \le \frac{c \log^\alpha(nd)}{\delta^2} \min\left(\frac{b^2}{n} + \frac{d}{n}, \frac{b}{\sqrt{n}}\left(1 + \frac{b}{\sqrt{n}}\right)\right).$$

- Small $d$ regime: $d/n$.
- Large $d$ regime: $b/\sqrt{n}$.

## Main Results: Excess Risk

For $\ell_1$-constrained least squares, with probability $1 - \delta$,

$$P\ell_{\hat{\beta}} - P\ell_{\beta^*} \leq \frac{c \log^{\alpha}(nd)}{\delta^2} \min\left(\frac{b^2}{n} + \frac{d}{n}, \frac{b}{\sqrt{n}}\left(1 + \frac{b}{\sqrt{n}}\right)\right).$$

Conditions:

1. $PY^2$ is bounded by a constant.
2. - $\|X\|_\infty$ bounded a.s.,
   - $X$ log concave and $\max_j \|\langle X, e_j \rangle\|_{L_2} \leq c$, or
   - $X$ log concave and isotropic.

# Application: Persistence

Consider $\ell_1$-constrained least squares,

$$\hat{\beta} = \arg \min_{\|\beta\|_{\ell_1^d} \leq b} P_n \ell_\beta.$$

Suppose that $PY^2$ is bounded by a constant and tails of $X$ decay nicely (e.g., $\|X\|_\infty$ bounded a.s. or $X$ log concave and isotropic).
Then for increasing $d_n$ and

$$b_n = o\left(\frac{\sqrt{n}}{\log^{3/2} n \log^{3/2}(nd_n)}\right),$$

$\ell_1$-constrained least squares is persistent
(i.e., $P\ell_{\hat{\beta}} - P\ell_{\beta^*} \to 0$).

# Application: Persistence

If $PY^2$ is bounded and tails of $X$ decay nicely, then $\ell_1$-constrained least squares is persistent provided that $d_n$ is increasing and

$$b_n = o\left(\frac{\sqrt{n}}{\log^{3/2} n \log^{3/2}(nd_n)}\right).$$

Previous Results (Greenshtein and Ritov, 2004):

1. $b_n = \omega(n^{1/2}/\log^{1/2} n)$ implies empirical minimization is not persistent for Gaussian $(X, Y)$.
2. $b_n = o(n^{1/2}/\log^{1/2} n)$ implies empirical minimization is persistent for Gaussian $(X, Y)$.
3. $b_n = o(n^{1/4}/\log^{1/4} n)$ implies empirical minimization is persistent under tail conditions on $(X, Y)$.

# Application: Convex Aggregation

Consider $b = 1$, so that the $\ell_1$-ball of radius $b$ is the convex hull of a dictionary of $d$ functions (the components of $X$).
Tsybakov (2003) showed that, for any aggregation scheme $\hat{\beta}$, the rate of convex aggregation satisfies

$$P\ell_{\hat{\beta}} - P\ell_{\beta^*} = \Omega\left(\min\left(\frac{d}{n}, \sqrt{\frac{\log d}{n}}\right)\right).$$

For bounded, isotropic distributions, our result implies that this rate can be achieved, up to log factors, by least squares over the convex hull of the dictionary.
Previous positive results (Tsybakov, 2003; Bunea, Tsybakov and Wegkamp, 2006) involved complicated estimators.

# Main Results: Oracle Inequality

For $\ell_1$-regularized least squares,

$$\hat{\beta} = \arg\min_{\beta} \left( P_n \ell_\beta + \rho_n \|\beta\|_{\ell_1^{d_n}} \right),$$

if

- $d_n$ increases but $\log d_n = o(n)$, and
- $Y$ and $\|X\|_\infty$ bounded a.s.,

then with probability at least $1 - o(1)$,

$$P\ell_{\hat{\beta}} \leq \inf_{\beta} \left( P\ell_\beta + c\rho_n \left( 1 + \|\beta\|_{\ell_1^{d_n}} \right) \right),$$

where

$$\rho_n = \frac{c \log^{3/2} n \log^{1/2}(d_n n)}{\sqrt{n}}.$$

# Outline of Talk

1. For $\ell_1$-constrained least squares, bounds on $P\ell_{\hat{\beta}} - P\ell_{\beta^*}$.
   - **Persistence:**
     For what $d_n, b_n \to \infty$ does $P\ell_{\hat{\beta}} - P\ell_{\beta^*} \to 0$?
   - **Convex Aggregation:**
     For $b = 1$ (convex combinations of dictionary functions), what is rate of $P\ell_{\hat{\beta}} - P\ell_{\beta^*}$?
2. For $\ell_1$-regularized least squares, oracle inequalities.
3. Proof ideas.

## Proof Ideas: 1. $\epsilon$-equivalence of $P$ and $P_n$ structures

Define

$$G_\lambda = \left\{ \frac{\lambda}{P(\ell_\beta - \ell_{\beta^*})}(\ell_\beta - \ell_{\beta^*}) : P(\ell_\beta - \ell_{\beta^*}) \geq \lambda \right\}.$$

Then:
$\mathbf{E} \sup_{g \in G_\lambda} |P_n g - Pg|$ is small
$\Rightarrow$ with high probability, for all $\beta$ with $P(\ell_\beta - \ell_{\beta^*}) \geq \lambda$,

$$(1-\epsilon)P(\ell_\beta - \ell_{\beta^*}) \leq P_n(\ell_\beta - \ell_{\beta^*}) \leq (1+\epsilon)P(\ell_\beta - \ell_{\beta^*})$$

$\Rightarrow P(\ell_{\hat\beta} - \ell_{\beta^*}) \leq \lambda$, where $\hat\beta = \arg\min_\beta P_n \ell_\beta$.

## Proof Ideas: 2. Symmetrization, subgaussian tails

For a class of functions $F$,

$$\mathbf{E} \sup_{f \in F} |P_n f - Pf| \leq \frac{2}{n} \mathbf{E} \sup_{f \in F} \left| \sum_i \epsilon_i f(X_i) \right|,$$

where $\epsilon_i$ are independent Rademacher ($\pm 1$) random variables.
(Giné and Zinn, 1984)
The Rademacher process

$$Z_\beta = \sum_i \epsilon_i \left( \ell_\beta(X_i, Y_i) - \ell_{\beta^*}(X_i, Y_i) \right)$$

indexed by $\beta$ in

$$T_\lambda = \left\{ \beta \in \mathbb{R}^d : \|\beta\|_{\ell_1^d} \leq b,\ P\ell_\beta - P\ell_{\beta^*} \leq \lambda \right\}$$

is **subgaussian** wrt a certain metric $d$. That is,
for every $\beta_1, \beta_2 \in T_\lambda$ and $t \geq 1$,

$$\Pr\left( |Z_{\beta_1} - Z_{\beta_2}| \geq t d(\beta_1, \beta_2) \right) \leq 2 \exp(-t^2/2).$$

## Proof Ideas: 2. Symmetrization, subgaussian tails

The Rademacher process

$$Z_\beta = \sum_i \epsilon_i \left(\ell_\beta(X_i, Y_i) - \ell_{\beta^*}(X_i, Y_i)\right)$$

indexed by $\beta$ in

$$T_\lambda = \left\{\beta \in \mathbb{R}^d : \|\beta\|_{\ell_1^d} \leq b,\ P\ell_\beta - P\ell_{\beta^*} \leq \lambda\right\}$$

is **subgaussian** wrt the metric

$$d(\beta_1, \beta_2) = 4 \max_i |\langle X_i, \beta_1 - \beta_2\rangle| \times \sup_{v \in \sqrt{\lambda} D \cap 2T} \left(\sum_i \langle X_i, v\rangle^2 + \sum_i \ell_{\beta^*}(X_i, Y_i)\right).$$

This is a random $\ell_\infty$ distance, scaled by the random $\ell_2$ diameter of $\sqrt{\lambda} D \cap 2T$.
Here, $D = \{x \in \mathbb{R}^d : P|\langle X, x\rangle|^2 \leq 1\}$.

# Proof Ideas: 3. Chaining

For a subgaussian process $\{Z_t\}$ indexed by a metric space $(T, d)$, and for $t_0 \in T$,

$$\mathbf{E} \sup_{t \in T} |Z_t - Z_{t_0}| \leq c\mathcal{D}(T, d) = c \int_0^{\mathrm{diam}(T,d)} \sqrt{\log N(\epsilon, T, d)}\, d\epsilon,$$

where $N(\epsilon, T, d)$ is the $\epsilon$ covering number of $T$.

## Proof Ideas: 4. Bounding the Entropy Integral

It suffices to calculate the entropy integral $\mathcal{D}(\sqrt{\lambda}D \cap 2bB_1^d, d)$. We can approximate this by

$$\mathcal{D}(\sqrt{\lambda}D \cap 2bB_1^d, d) \leq \min\left(\mathcal{D}(\sqrt{\lambda}D, d), \mathcal{D}(2bB_1^d, d)\right).$$

This leads to:

$$P\ell_{\hat{\beta}} - P\ell_{\beta^*} \leq \frac{c\log^{\alpha}(nd)}{\delta^2}\min\left(\frac{b^2}{n} + \frac{d}{n}, \frac{b}{\sqrt{n}}\left(1 + \frac{b}{\sqrt{n}}\right)\right).$$

## Proof Ideas: 5. Oracle Inequalities

We get an isomorphic condition on $\{\ell_\beta - \ell_{\beta^*}\}$,

$$\frac{1}{2}P_n(\ell_\beta - \ell_{\beta^*}) - \epsilon_n \le P(\ell_\beta - \ell_{\beta^*}) \le 2P_n(\ell_\beta - \ell_{\beta^*}) + \epsilon_n,$$

and this implies that $\hat{\beta} = \arg\min_\beta (P_n\ell_\beta + c\epsilon_n)$ has

$$P\ell_\beta \le \inf_\beta \left(P\ell_\beta + c'\epsilon_n\right).$$

This leads to oracle inequality: For $\ell_1$-regularized least squares,

$$\hat{\beta} = \arg\min_\beta \left(P_n\ell_\beta + \rho_n\|\beta\|_{\ell_1^{d_n}}\right),$$

with probability at least $1 - o(1)$,

$$P\ell_{\hat{\beta}} \le \inf_\beta \left(P\ell_\beta + c\rho_n\left(1 + \|\beta\|_{\ell_1^{d_n}}\right)\right).$$

# Outline of Talk

1. For $\ell_1$-constrained least squares,

$$P\ell_{\hat{\beta}} - P\ell_{\beta^*} \leq \frac{c\log^{\alpha}(nd)}{\delta^2} \min\left(\frac{b^2}{n} + \frac{d}{n}, \frac{b}{\sqrt{n}}\left(1 + \frac{b}{\sqrt{n}}\right)\right).$$

   - **Persistence:**
     If $b_n = \tilde{o}(\sqrt{n})$, then $P\ell_{\hat{\beta}} - P\ell_{\beta^*} \to 0$.
   - **Convex Aggregation:**
     Empirical risk minimization gives optimal rate (up to log factors): $\tilde{O}\left(\min(d/n, \sqrt{\log d/n})\right)$.

2. For $\ell_1$-regularized least squares, oracle inequalities.
3. Proof ideas: subgaussian Rademacher process.