# Resilience Beyond Mean Estimation

Logistics: Pset 2 due today
Pset 3 should be out by tomorrow afternoon

---

## Recap:

- Defined general $G_\uparrow + G_\downarrow \to$ bounds on moduli of continuity
- Second moment estimation: higher moments $\Rightarrow$ resilience (nuclear norm reduction)
- Linear regression
  - Hypercontractivity + bounded noise

## Today:

- Finish linear regression
- State result for classification
- Pay back debts: prove missing dual norm results

---

Linear regression: $Y = \langle \theta^*, X \rangle + \xi$, $\mathbb{E}[X\xi] = 0$
$\Leftrightarrow \theta^*$ is least squares estimator for $(X, Y)$

2 conditions: $\forall\, v \in \mathbb{R}^d$

(1) $\mathbb{E}_p[\langle X, v \rangle^4] \le \kappa\, \mathbb{E}_p[\langle X, v \rangle^2]^2$ (hypercontractivity)

(2) $\mathbb{E}_p[X \xi X^T] \le \sigma^2 \cdot \mathbb{E}_p[XX^T]$ (bounded noise)

**Proposition.** If (1) and (2) hold, then $p$ is $(\rho, 5\rho, \varepsilon)$-resistant for $\rho = 2\sigma^2\varepsilon$ as long as $\varepsilon \le \frac{1}{8}$ and $\varepsilon(K-1) \le \frac{1}{6}$.

**Interpretation:** $L(p, \Theta) = \mathbb{E}_p[(Y - \langle \Theta, X \rangle)^2] - \mathbb{E}_p[(Y - \langle \theta^*, X \rangle)^2]$ ($Z^2$)

squared

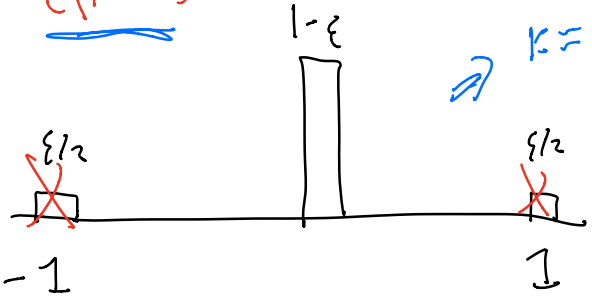
Resilience $\Rightarrow$ can robustly estimate w/ loss $\le 5\rho$

$= 10\sigma^2\varepsilon$. ← roughly same sort of result as for mean estimation

→ Definitely need $\varepsilon \le \frac{1}{2}$.

$K \ge 1$ $\quad \varepsilon(K-1) \to$ Need, $K = \Theta\left(\frac{1}{\varepsilon}\right)$

$K \asymp \frac{1}{\varepsilon}$ in this setting




---

**Proof.** Two conditions, $G_\downarrow$ and $G_\uparrow$

$G_\downarrow$: $L(r, \theta^*) \le \rho$, whenever $r \le \frac{\rho}{1-\varepsilon}$

$G_\uparrow$: If $L(r, \theta) \le \rho$, then $L(p, \theta) \le 5\rho$

→ $L(r, \theta) = (\theta - \theta^*(r))^T S_r (\theta - \theta^*(r))$

$S_r = \mathbb{E}_r[XX^T]$

$G\downarrow$: $(\theta^*(p) - \theta^*(r))^T S_r (\theta^*(p) - \theta^*(r)) \le \rho$

$\hookrightarrow \|S_r^{1/2}(\theta^*(r) - \theta^*(p))\|_2$

$G\uparrow$: If $(\theta - \theta^*(r))^T S_r (\theta - \theta^*(r)) \le \rho$,

then $(\theta - \theta^*(p))^T S_p (\theta - \theta^*(r)) \le 5\rho$.

Proof strategy:

(1) $S_r \approx S_p \rightarrow \frac{1}{2} S_p \le S_r \le \frac{3}{2} S_p$ ← hypercontractivity

(2) $\|S_p^{1/2}(\theta^*(r) - \theta^*(p))\|_2$ small by resilience ← bounded noise

(1)

$S_p = \mathbb{E}_p[XX^T]$

$S_r = \mathbb{E}_r[XX^T]$

For all $v$, $\frac{1}{2} v^T S_p v \le v^T S_r v$

$\frac{1}{2}\mathbb{E}_p[\langle X, v\rangle^2] \le \mathbb{E}_r[\langle X, v\rangle^2]$

Suffices to show: $\left|\mathbb{E}_r[\langle X, v\rangle^2] - \mathbb{E}_p[\langle X, v\rangle^2]\right| \le \frac{1}{2}\mathbb{E}_p[\langle X, v\rangle^2]$

mean of $\langle X, v\rangle^2$ under $r$ — mean under $p$

Recall Chebyshev: If $\mathrm{Var}[Z] \le \sigma^2$, then $Z$ is $\left(\varepsilon, \frac{\sigma\sqrt{\varepsilon}}{1-\varepsilon}\right)$-resilient.

$Z = \langle X, v \rangle^2$

$\mathrm{Var}[Z] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2 = \mathbb{E}[\langle X, v \rangle^4] - \mathbb{E}[\langle X, v \rangle^2]^2 \le (k-1)\,\mathbb{E}[\langle X, v \rangle^2]^2$

$\Rightarrow$ Can take $\sigma = \sqrt{k} \cdot \mathbb{E}[\langle X, v \rangle^2]$

$$\left| \mathbb{E}_r[\langle X, v \rangle^2] - \mathbb{E}_p[\langle X, v \rangle^2] \right| \le 2 \cdot \frac{\sigma\sqrt{\varepsilon}}{1-\varepsilon} = \frac{2\sqrt{\varepsilon(k-1)}}{1-\varepsilon} \cdot \mathbb{E}[\langle X, v \rangle^2]$$

Need this to be $\le \frac{1}{2}$.

Holds if $\varepsilon \le \frac{1}{8}$ and $\varepsilon(k-1) \le \frac{1}{6}$.

$$\Rightarrow \left| \mathbb{E}_r[\langle X, v \rangle^2] - \mathbb{E}_p[\langle X, v \rangle^2] \right| \le \frac{1}{2}\,\mathbb{E}[\langle X, v \rangle^2]$$

$$\Rightarrow S_r \ge \frac{1}{2} S_p$$

$$\Rightarrow S_r \le \frac{3}{2} S_p \,. \quad \blacksquare$$

---

Claim.

Want to somehow show $\theta^*(p) - \theta^*(r)$ is small

Claim. $\theta^*(r) - \theta^*(p) = S_r^{-1}\,\mathbb{E}_r[XZ]$

$$S_r^{-1}\,\mathbb{E}_r[XZ] = \mathbb{E}_r[XX^T]^{-1}\,\mathbb{E}_r[XZ]$$

$Z = Y - \langle \theta^*(p), X \rangle$

$$= \mathbb{E}_r[XX^T]^{-1}\mathbb{E}_r[XY - XX^T\theta^*(p)]$$

$$= \mathbb{E}_r[XX^T]^{-1}\mathbb{E}_r[XY] - \theta^*(p)$$ (cancel)

$$= \theta^*(r) - \theta^*(p).$$

least squares formula

$$\hat{\theta}_{ols} = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n X_i Y_i\right)$$

$$\mathbb{E}_p[XZ] = 0 \quad \text{by definition}$$

$$\theta^*(p) - \theta^*(r) = S_r^{-1}\left(\mathbb{E}_p[XZ] - \mathbb{E}_r[XZ]\right)$$

resilience (bounded noise)

---

$$(\theta^*(p) - \theta^*(r))^T S_r (\theta^*(p) - \theta^*(r)) \le \rho$$

$\hookrightarrow \|S_r^{1/2}(\theta^*(r) - \theta^*(p))\|_2$

$$\left(\mathbb{E}_p[XZ] - \mathbb{E}_r[XZ]\right)^T S_r^{-1} \cdot S_r \cdot S_r^{-1}\left(\mathbb{E}_p[XZ] - \mathbb{E}_r[XZ]\right)$$

$$= \left(\mathbb{E}_p[XZ] - \mathbb{E}_r[XZ]\right)^T S_r^{-1}\left(\mathbb{E}_p[XZ] - \mathbb{E}_r[XZ]\right)$$

$$\le 2\cdot\left(\mathbb{E}_p[XZ] - \mathbb{E}_r[XZ]\right)^T S_p^{-1}\left(\mathbb{E}_p[XZ] - \mathbb{E}_r[XZ]\right)$$

$$= 2\|S_p^{-1/2}\left(\mathbb{E}_p[XZ] - \mathbb{E}_r[XZ]\right)\|_2$$

need to bound this $\rightarrow$ $= \mathbb{E}_p[W] - \mathbb{E}_r[W]$

Reduces to resilient for variable $W = \Sigma_p^{-1/2} X Z$

$$\mathrm{Var}_p[W] = \mathbb{E}_p[WW^T]$$
$$= \mathbb{E}_p[\Sigma_p^{-1/2} X Z^2 X^T \Sigma_p^{-1/2}]$$
$$= \Sigma_p^{-1/2}\, \mathbb{E}_p[XZ^2X^T]\, \Sigma_p^{-1/2}$$
$$\leq \sigma^2 \cdot \mathbb{E}_p[XX^T]$$
$$= \sigma^2 \cdot \Sigma_p$$
$$\leq \sigma^2 \cdot \Sigma_p^{-1/2} \Sigma_p \Sigma_p^{-1/2} = \sigma^2 \cdot I$$

$= \mathcal{O}(\sigma^2 \cdot \zeta)$

$\Rightarrow G\downarrow$

$\Rightarrow$ bounded variance $\Rightarrow$ resilience

---

$G\downarrow$: $(\theta^*(p) - \theta^*(r))^T \Sigma_r (\theta^*(p) - \theta^*(r)) \leq \rho$

$\hookrightarrow \|\Sigma_r^{1/2}(\theta^*(r) - \theta^*(p))\|_2$

$\Sigma\uparrow$: If $(\theta - \theta^*(r))^T \Sigma_r (\theta - \theta^*(r)) \leq \rho$,

then $(\theta - \theta^*(p))^T \Sigma_p (\theta - \theta^*(p)) \leq 5\rho$.

$$\sqrt{(\theta - \theta^*(p))^T \Sigma_p (\theta - \theta^*(p))}$$
$$\leq \sqrt{(\theta - \theta^*(r))^T \Sigma_p (\theta - \theta^*(r))} + \sqrt{(\theta^*(r) - \theta^*(p))^T \Sigma_p (\theta^*(r) - \theta^*(p))}$$

$\Sigma_r + G\uparrow$ $\qquad$ $\Sigma_r + G\downarrow$

$\leq \mathcal{O}(\sigma^2 \cdot \zeta)$.

∎

$$\nabla_\theta [XZ] = \nabla_\theta [X(Y - \langle \theta^*, X \rangle)]$$
$$= \nabla_\theta [X] - \nabla_\theta [XX^\top] \theta^* = 0$$

~~$\nabla_\theta [Z|X] = 0 \quad \forall X$ ⇒ assume linear model~~
~~, misspecification~~

---

$(X, Y)$

Append features?

$(X + H, Y)$

$H \sim N(0, \lambda \cdot I)$

$\Updownarrow$

ridge regression w/ params $\lambda$

$X$ hypercontractive ⇒ $X + H$

---

Linear regression

Recap: hypercont. $\Rightarrow \frac{1}{2} S_p \le S_r \le \frac{3}{2} S_p$

resilience of $S_p^{-1/2} X \xi \Rightarrow$ bounded noise

$\Rightarrow$ combine w/ triangle ineq.

---

## Linear classification

$(X, Y) \in \mathbb{R}^d \times \{-1, +1\}$

linear classifier — 0/1-loss

$$L(p, \theta) = \mathbb{P}_p[Y \neq \operatorname{sign}(\langle X, \theta \rangle)]$$

hinge loss

$$B(p, \theta) = \mathbb{E}_{x,y}[\max(1 - y\langle \theta, x \rangle, 0)]$$

hinge $\ge$ 0/1

$\hookrightarrow G_\downarrow / G_\uparrow$

"bridging"

---

**Prop.** Suppose $p$ satisfies:

(1) $\mathbb{E}_{(x,y)\sim p}[\max(1 - y\langle \theta^*, x \rangle, 0)] \le (1-\varepsilon)\beta_1$

(2) If $\mathbb{P}_{(x,y)\sim p}\left[y\langle \theta, x \rangle \le \frac{1}{2}\right] \le \varepsilon + 2(1-\varepsilon)\beta_1$

then $\mathbb{P}_{(x,y)}[y\langle \theta, x \rangle \le 0] \le \beta_2$.

Then $p$ is $(\rho_1, \rho_2, \varepsilon)$-resistant for $L$ and $B$.

---

Dual norms.

Schatten $p$-norms

$X$, singular values $\sigma_1, \dots, \sigma_n$

$$\|X\|_{S(p)} = \sqrt[p]{\sigma_1^p + \dots + \sigma_n^p}$$

$S(1)$: nuclear norm

$S(\infty)$: operator norm

Thm. $S(p)$ and $S(q)$ are dual norms if $\frac{1}{p} + \frac{1}{q} = 1$.

Key lemma.

For any $A, B$ w/ singular values $\sigma, \tau$ (in sorted order),

$$\langle A, B \rangle \le \sum_{i=1}^{n} \sigma_i \tau_i.$$

---

1) Reduce to PSD

2) Prove for PSD

① Lemma.

$$\langle A, B\rangle^2 \le \langle (A^TA)^{1/2}, (B^TB)^{1/2}\rangle \langle (AA^T)^{1/2}, (BB^T)^{1/2}\rangle$$

$$\begin{bmatrix} \lambda (A^TA)^{1/2} & A \\ A^T & \frac{1}{\lambda}(AA^T)^{1/2} \end{bmatrix} \succeq 0$$

$$\begin{bmatrix} \lambda (B^TB)^{1/2} & -B \\ -B^T & \frac{1}{\lambda}(BB^T)^{1/2} \end{bmatrix} \succeq 0$$

$\Rightarrow$ take matrix dot product

(analogous to Cauchy-Schwarz)

② Lemma.

$$\sum_{i=1}^{k} B_{ii} \le \sum_{i=1}^{k} \tau_i$$

$\leftarrow$ Cauchy interlacing theorem

$$\langle A, B\rangle = \langle \operatorname{diag}(\sigma_1, \dots, \sigma_n), B\rangle$$

$$\sum_{i=1}^{n} \sigma_i B_{ii} \le \sum_{i=1}^{n} \sigma_i \tau_i$$

"stochastic dominance"

(Abel summation)