On high-dimensional robust regression

How to pick the loss in high-dimensional regression?

Noureddine El Karoui
joint with Bean, Bickel, Lim, Yu

Department of Statistics
UC, Berkeley

Stanford MMDS
July 10th, 2012

Basic question

Consider linear regression model:

$$Y_i = X_i' \beta_0 + \epsilon_i \ , i = 1, \ldots, n \ .$$

Here $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$, $\beta_0 \in \mathbb{R}^p$ and $\epsilon_i \in \mathbb{R}$.

- Aim: estimate β_0.
- Setting: X_i's vectors of predictors. ϵ_i's noise.
- Standard method: (say $p < n$): pick $\widehat{\beta}$ as

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(Y_i - X_i' \beta) \ , \text{ where } \rho \text{ is a function.}$$

Question: how to pick ρ?

How to pick ρ?

Very classical question.
Much work on this starting with Fisher in 30's.
Very nice work in the late 60's, 70's, 80's on properties of these estimators.
Contributors include: Relles, Huber ('72), Portnoy ('84-85), Mammen ('91), Yohai, Bickel, etc...
Short answer: in low dimension, if f_ϵ is density of ϵ, ϵ i.i.d, pick

$$\boxed{\rho = -\log f_\epsilon \ .}$$

Remarkable fact: independent of design matrix, X.

An example: double exponential errors

ϵ_i's double exponential, i.e $f_\epsilon(x) = \exp(-|x|)/2$.
According to classical results/intuition, l_1 should be optimal

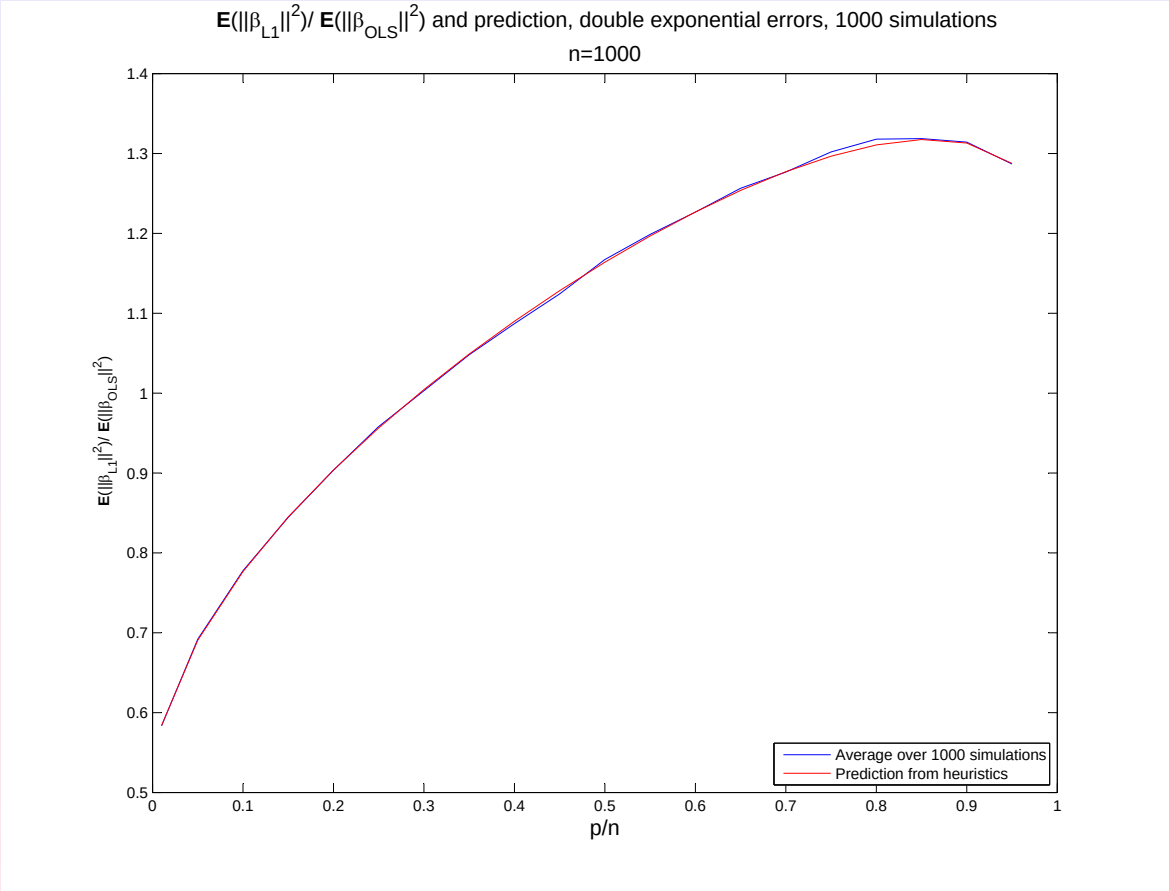

Figure: $\mathbf{E}\left(\|\widehat{\beta}_{l_1} - \beta_0\|^2\right) / \mathbf{E}\left(\|\widehat{\beta}_{OLS} - \beta_0\|^2\right)$

A proposition for ρ

Let $p_2(x) = x^2/2$. Suppose ϵ has log-concave density, f_ϵ. For sake of argument, assume f_ϵ known.
For reasons explained later, let us try

$$\boxed{\rho_{opt} = (p_2 + r_{opt}^2 \log \phi_{r_{opt}} \star f_\epsilon)^* - p_2 \ .}$$

$$\text{where } r_{opt} = \min\{r : r^2 I_\epsilon(r) = p/n\} \ .$$

ϕ_r: gaussian density with variance r^2.
$I_\epsilon(r)$: Fisher information of $\phi_r \star f_\epsilon$
$g^*(x) = \sup_y(xy - g(y))$, Fenchel-Legendre dual of g

Comparison ρ_{opt} to ℓ_1, double exponential errors

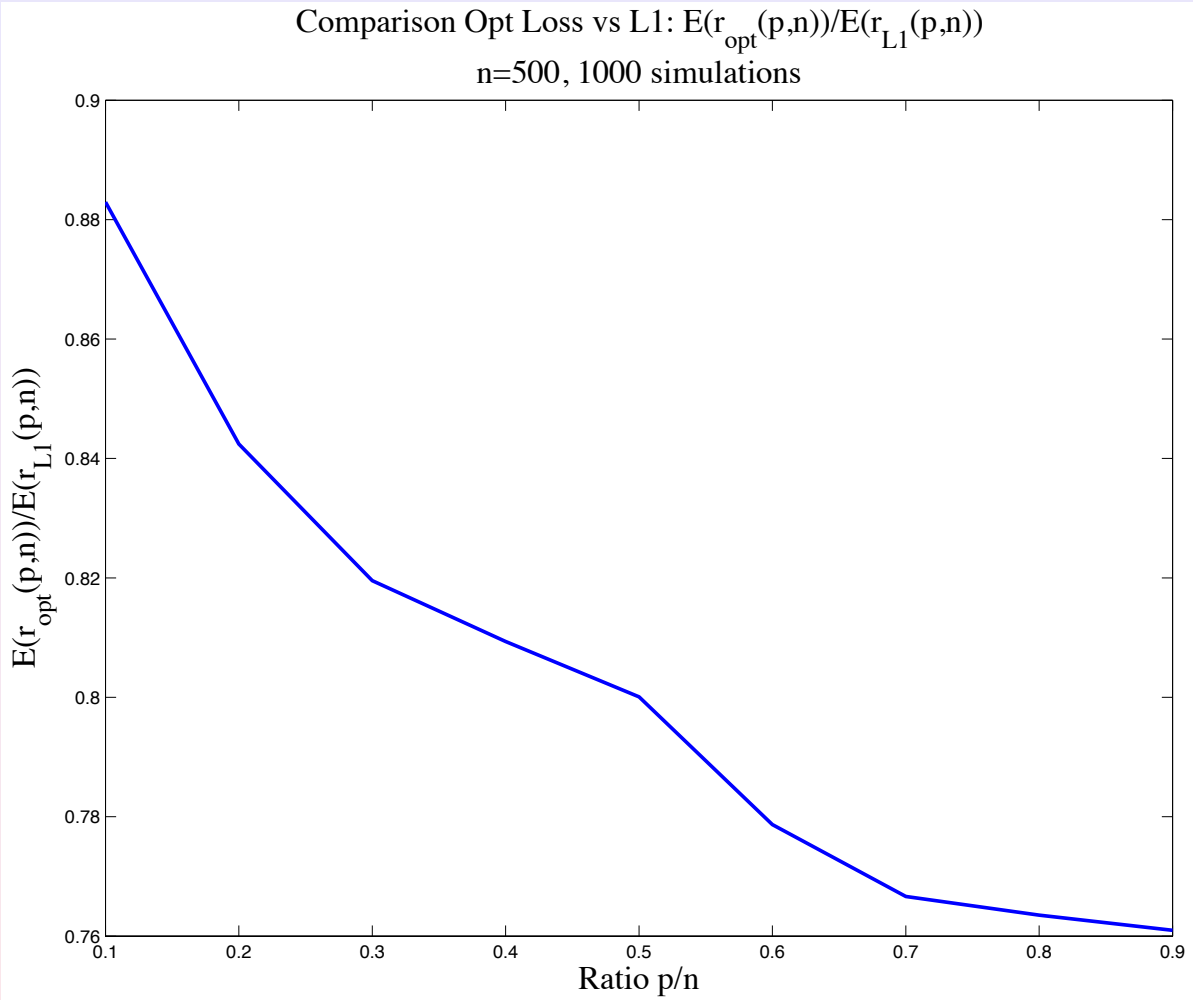

Figure: $\mathbf{E}\left(\|\widehat{\beta}_{opt} - \beta_0\|^2\right) / \mathbf{E}\left(\|\widehat{\beta}_{l_1} - \beta_0\|^2\right)$, double exponential errors.

Comparison ρ_{opt} to ℓ_2, double exponential errors

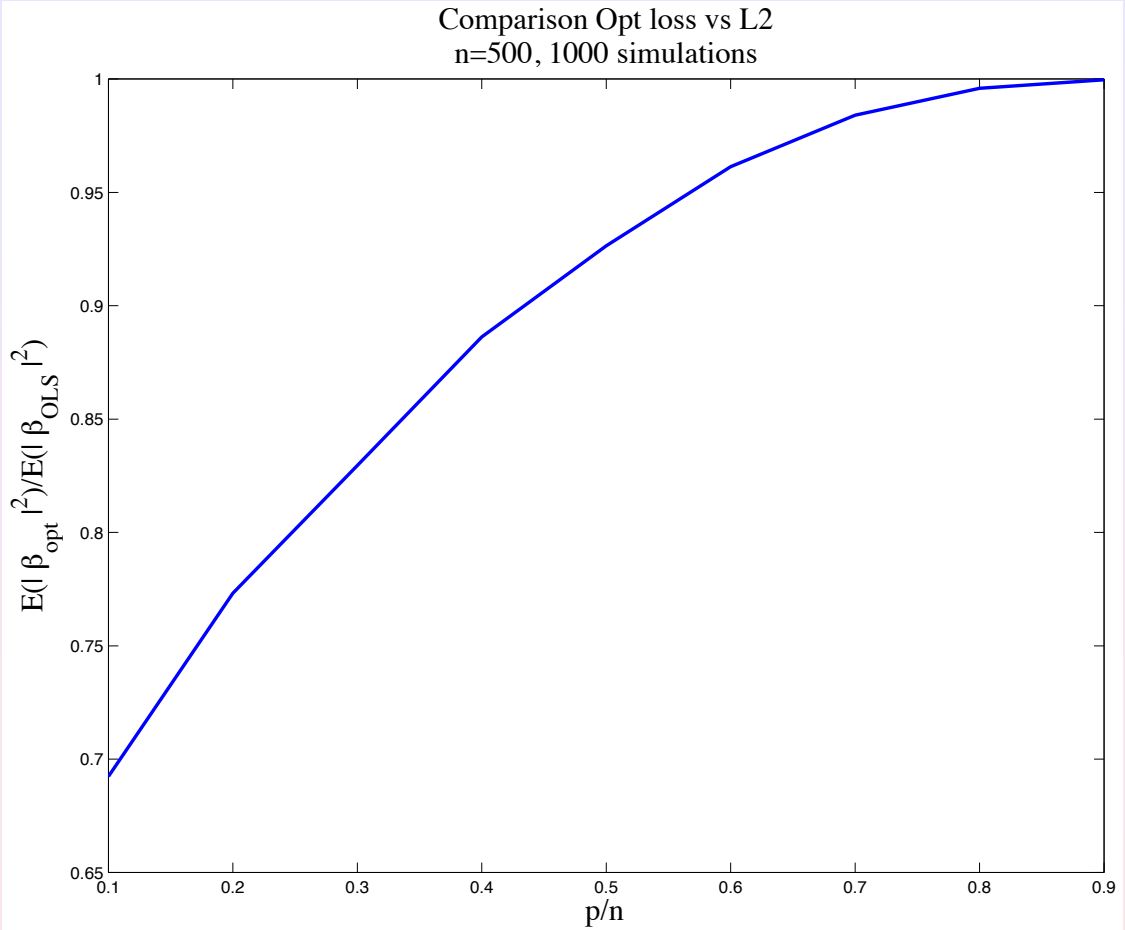

Figure: $\mathbf{E}\left(\|\widehat{\beta}_{opt}-\beta_0\|^2\right)/\mathbf{E}\left(\|\widehat{\beta}_{OLS}-\beta_0\|^2\right)$, double exponential errors.

Aim of talk

- Understand these pictures/phenomena
- Caveat: optimality now sensitive to design. Will get back to key properties of design
- Also: bootstrap appears problematic in this context
- Many interesting statistical phenomena at play

Why work under p/n not close to 0?

Plan

1. Computation of risk of robust regression estimators
2. Inferential questions
3. Optimization with respect to loss function
4. Penalized case: risk computation and optimality in the ℓ_2-penalized case

Characterization of solution of robust regression problem

Suppose $p/n \to \kappa \in (0,1)$. Temporarily, $X_i \overset{iid}{\backsim} \mathcal{N}(0, \mathrm{Id}_p)$.

Proposition

Under regularity conditions on $\{\epsilon_i\}$ *and* ρ, $\|\widehat{\beta} - \beta_0\|$ *is asymptotically deterministic. Call* $r_\rho(\kappa)$ *its limit and* $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa)Z$, *where* $Z \sim \mathcal{N}(0,1)$, *independent of* ϵ. *For a* c *deterministic, we have*

$$\begin{cases} \mathbf{E}\left(prox_c(\rho)(\hat{z}_\epsilon)\right) & = 1 - \kappa \ , \\ \kappa r_\rho^2(\kappa) & = \mathbf{E}\left([\hat{z}_\epsilon - prox_c(\rho)(\hat{z}_\epsilon)]^2\right) \ . \end{cases}$$

By definition, (Moreau '65), for convex function f,

$$\mathrm{prox}_c(f)(x) = \mathrm{argmin}_y \left(f(y) + \frac{1}{2c}(x-y)^2 \right) \ .$$

Much more can be said: elliptical models, heteroskedastic ϵ_i's, weighted robust regression, no need for normality of X_i etc... Approach can handle penalized versions.

On the residuals

Call $R_i = Y_i - \widehat{\beta}' X_i$, the i-th residual. In the asymptotic limit,

$$\boxed{R_i \overset{\mathcal{L}}{=} \mathrm{prox}_c(\rho)(\epsilon_i + r_\rho(\kappa) Z_i)}$$

where $Z_i \sim \mathcal{N}(0,1)$ independent of ϵ_i.
Somewhat complicated relationship between ρ, distribution of ϵ_i and distribution of R_i. Very different from classical setting of p/n close to 0.

Suppose $X_i = \lambda_i \mathcal{X}_i$, where $\mathcal{X}_i$ is $\mathcal{N}(0, \mathrm{Id}_p)$, λ_i random variables independent of $\mathcal{X}_i$.
$\|\widehat{\beta} - \beta_0\|$ still asymptotically deterministic, limit denoted by $r_\rho(\kappa)$.

Proposition

Let us now call $\hat{z}_\epsilon(i) = \epsilon_i + \lambda_i r_\rho(\kappa) Z_i$, where $Z_i \sim \mathcal{N}(0,1)$ are i.i.d and independent of $\{\epsilon\}_{i=1}^n$ and $\{\lambda_i\}_{i=1}^n$. We can determine $r_\rho(\kappa)$ through solving

$$\begin{cases} \lim_{n\to\infty} \sum_{i=1}^n \dfrac{\mathbf{E}\left(\left[prox_{c\lambda_i^2}(\rho)\right]'(\hat{z}_\epsilon(i))\right)}{n} & = 1-\kappa\,, \\ \lim_{n\to\infty} \sum_{i=1}^n \dfrac{\mathbf{E}\left(\lambda_i^{-2}[\hat{z}_\epsilon(i) - prox_{c\lambda_i^2}(\rho)(\hat{z}_\epsilon(i))]^2\right)}{n} & = \kappa r_\rho^2(\kappa)\,, \end{cases} \quad \text{(S1)}$$

where c again positive deterministic constant determined from above system.

Key elements of previous results

How proof and heuristics work?
Key elements

- concentration of quadratic forms in $\mathcal{X}_i$; consequence: geometry of dataset influences crucially result. No universality.
- leave-one-out ideas.
- martingale ideas
- connection with ideas in random matrix theory and convex analysis

Surprise, in particular in connection to ℓ_1 regression: it is a random matrix problem!

Normal design: invariance remarks

When X_i are i.i.d $\mathcal{N}(0, \Sigma)$, then

$$\widehat{\beta}(\rho; \beta_0, \Sigma) \stackrel{\mathcal{L}}{=} \beta_0 + \|\widehat{\beta}(\rho; 0, \mathrm{Id}_p)\| \Sigma^{-1/2} u \ ,$$

where u unif of sphere of radius 1 in $\mathbb{R}^p$, **independent** of $\|\widehat{\beta}(\rho; 0, \mathrm{Id}_p)\|$. Consequences:

- easy to handle case $\beta_0 \neq 0$ and $\Sigma \neq \mathrm{Id}_p$.
- elliptical setting works similarly
- can do inference on $v'\beta_0$, any v given. (Surprise to experts.)
- not complicated to include intercept (several manners to deal with that)
- side note: bootstrap

Fact: quality of inference depends only on $\|\widehat{\beta}(\rho; 0, \mathrm{Id}_p)\|$; its limit $r_\rho(\kappa)$ characterized before.
Natural to optimize $r_\rho(\kappa)$ over ρ

Optimality in any ℓ_q loss

Suppose wish to measure quality of estimator as

$$\mathbf{E}\left(\|\widehat{\beta}-\beta_0\|_q\right) \ , q \neq 2 \text{ for instance.}$$

Stochastic representation yields immediately

$$\mathbf{E}\left(\|\widehat{\beta}-\beta_0\|_q\right) = \mathbf{E}\left(\|\widehat{\beta}(\rho;0,\mathrm{Id}_p)\|_2\right)\mathbf{E}\left(\|\Sigma^{-1/2}u\|_q\right) \ .$$

Hence optimizing $r_\rho(\kappa)$ yields asymptotically optimal performance in any ℓ_q norm, not only ℓ_2.

Optimizing $r_\rho(\kappa)$ over ρ

Recall system: if $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa) Z$, with $Z \sim \mathcal{N}(0,1)$,

$$\left\{ \begin{array}{cc} \mathbf{E}\left(\mathrm{prox}_{\mathbf{c}}(\rho)(\hat{z}_\epsilon)\right) & = 1-\kappa \ , \\ \kappa r_\rho^2(\kappa) & = \mathbf{E}\left([\hat{z}_\epsilon - \mathrm{prox}_{\mathbf{c}}(\rho)(\hat{z}_\epsilon)]^2 \right) \ . \end{array} \right.$$

Possible to optimize $r_\rho(\kappa)$ over ρ!

Strategy

1. Write problem as feasibility problem in r
2. Use Moreau's fundamental prox-identity:

$$x = \mathrm{prox}_1(\rho)(x) + \mathrm{prox}_1(\rho^*)(x) \ .$$

   to rewrite system. Natural variable: $\mathrm{prox}_1(\rho^*)$
3. Cauchy-Schwarz yields lower bound on possible values of $r^2 I_\epsilon(r)$, where $I_\epsilon(r)$ is Fisher information of $\epsilon + rZ$
4. Come up with good $\mathrm{prox}_1(\rho^*)$ for which lower bound is achieved.
5. Go from optimal $\mathrm{prox}_1(\rho^*)$ to optimal ρ

Following this strategy, we get that, if $p_2(x) = x^2/2$, if $-\log f_\epsilon$ convex,

$$\boxed{\rho_{opt} = (p_2 + r_{opt}^2 \log \phi_{r_{opt}} \star f_\epsilon)^* - p_2 \ .}$$

$$\text{where } r_{opt} = \min\{r : r^2 I_\epsilon(r) = p/n\} \ .$$

Further remarks on optimal loss

- For Gaussian errors, ℓ_2 still optimal
- As $p/n \to 1$, performance of ℓ_2 becomes optimal
- However, limit of optimal loss not ℓ_2
- Also, ρ_{opt} proposed above convex

Plot for $p/n = .5$, double exponential errors

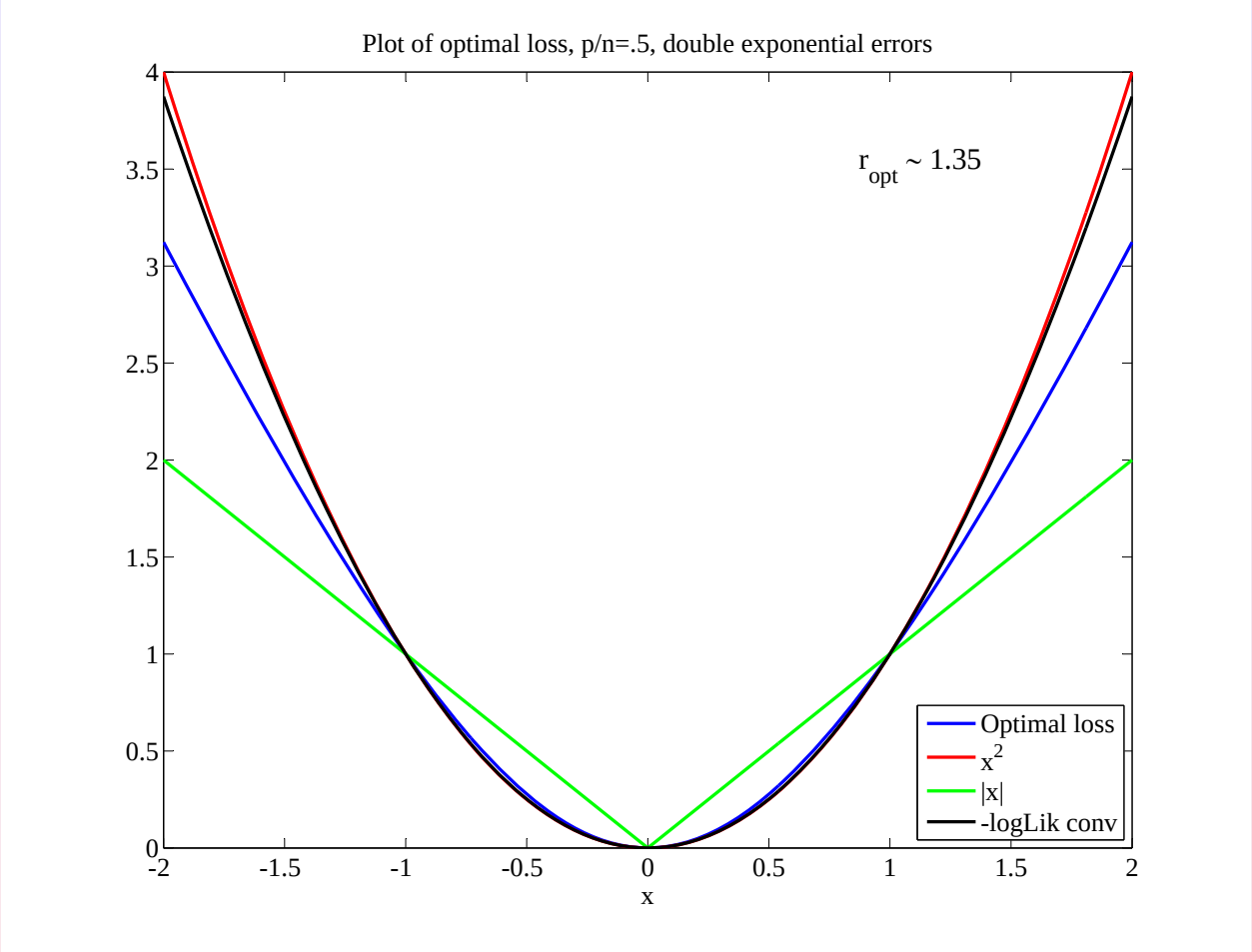

Figure: Some "natural" objective functions

On penalized regression

What about the case of penalized regression, i.e:

$$\widehat{\beta} = \text{argmin}_{\beta}\, \rho(Y_i - \beta' X_i) + \tau P(\beta)\ .$$

Can handle that, too. At this point, need

- $P(\beta) = \sum_{i=1}^{p} f_i(\beta_i)$,
- $\text{cov}\,(X_i) = \text{Id}_p$

Possible to characterize the limit $\widehat{\beta} - \beta_0$.
See also work on Lasso of Donoho-Maleki-Montanari, Bayati-Montanari. Approach is different.

System for $\|\widehat{\beta} - \beta_0\|$, elliptical setting

$\|\widehat{\beta} - \beta_0\|$ asymptotically deterministic.
Call $\widehat{\beta}_{(i)}$ leave-one-out estimate of β and
$\tilde{r}_{i,(i)} = \epsilon_i - (\widehat{\beta}_{(i)} - \beta_0)' X_i$. Below $\nu(\tau)$ and c_τ are unknowns. Call

$$Z_k \overset{\mathcal{L}}{=} \mathcal{N}\left(\beta_0(k), \frac{1}{n\nu(\tau)^2}\mathbf{E}\left(\frac{([\mathrm{prox}_{c_\tau \lambda_i^2}(\rho)](\tilde{r}_{i,(i)}) - \tilde{r}_{i,(i)})^2}{\lambda_i^2}\right)\right) .$$

We have asymptotically, when p/n has finite limit,

$$\begin{cases} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left([\mathrm{prox}_{c_\tau \lambda_i^2}(\rho)]'(\tilde{r}_{i,(i)})\right) & = 1 - \nu(\tau) , \\ \lim_{n\to\infty} \frac{p}{n}\frac{1}{p}\sum_{k=1}^p \mathbf{E}\left([\mathrm{prox}_{K_\tau}(f_k)]'(Z_k)\right) & = \nu(\tau) , \\ \forall 1 \leq k \leq p \, , \mathrm{prox}_{K_\tau}(f_k)\,[Z_k] & \overset{\mathcal{L}}{=} \widehat{\beta}_k . \end{cases}$$

with $K_\tau = \frac{\tau c_\tau}{n\nu(\tau)}$

Last p equations relate asymptotic value of $\|\widehat{\beta} - \beta_0\|^2$ to $\nu(\tau)$ and c_τ. Yields a system of 3 equations in three unknowns $\|\widehat{\beta} - \beta_0\|$, $\nu(\tau)$ and c_τ.

Optimization when $P(\beta) = \|\beta\|^2/2$

Suppose want to minimize $\|\widehat{\beta} - \beta_0\|_2$ for Tikhonov penalty.
Then optimal loss is again in family found above.
However, r_{opt} changes. Now it is

$$r_{opt} = \min \left\{ r : r^2 = \frac{\|\beta_0\|^2}{1 + \frac{n}{p} I_\epsilon(r) \|\beta_0\|^2} \right\} .$$

(Also, $\tau_{opt}/n = p/n - r_{opt}^2 I_\epsilon(r_{opt})$)

Conclusions

- Saw interplay between loss function and error distribution in high-dimensional robust regression
- Optimal loss computable in high-dimension
- It is **not** maximum-likelihood
- Inference is possible in our context
- Problems with the bootstrap (not touched in much details here)
- Can do penalized regression
- Optimal loss "canonical" as it is also optimal in the Tikhonov regularized context.