An Ideal Regression Method Satisfies:

Regularized M-estimator

\[
\tilde{f}_M \in \arg \min_{f \in \mathbb{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho(y_i - f(x_i)) + \frac{\lambda}{2} \| f \|_{F_2}^2 \right\},
\]
where \( \mathbb{H} \) is some reproducing kernel Hilbert space, \( \rho \) is a loss function, and \( \lambda \in \mathbb{R} \), is some loss function

"regularized empirical risk minimization"

Loss consistent

As \( n \to \infty \) and \( \lambda \to 0 \),

\[
\mathbb{E}(\rho(y - \tilde{f}_M(x))) \to \inf_{f \in \mathbb{H}} \mathbb{E}(\rho(y - f(x))).
\]

Under mild conditions on the kernel, distribution and loss, regularized M-estimators are consistent.

Robust

Bounded response: result remains in a bounded set, no matter how a single training pair is perturbed.

- A rather weak form of robustness, essentially requiring nonzero finite sample breakdown point.
- Most known robust M-estimators are based on minimizing a non-convex loss function.

Computationally tractable

The estimator can be found in polynomial time.

- For instance, when \( \rho \) is convex.

### Contributions

- No, for M-estimators (we prove formally)
- Yes, for the proposed algorithm

### Theoretical Properties

#### Robustness?

**Theorem 3:** Assume \( \ell \) is Lipschitz and \( \psi \) is bounded. Consider the perturbation of a single data point \((y_i, x_i)\). The convex relaxation of VM maintains bounded response if

1. Either \( \psi \) or \( (x_i, y_i) \) remains bounded.
2. \( \nabla \ell(y_i, y) \to \infty, \forall (x_i, y_i) \to \infty \) and \( \ell(\cdot, k(x_i, x_j)) \to \infty \).

#### Experiments

### Synthetic: Linear model with Gaussian noise.

\( d = 5, n_{train} = 100, n_{test} = 10000 \).

#### Methods:

- **blue**—non-convex, **red**—proposed

#### Performance Measure:

RMSE on clean test data.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Learning Curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2</td>
<td>4.99</td>
</tr>
<tr>
<td>Huber</td>
<td>5.16</td>
</tr>
</tbody>
</table>

#### Theoretical Properties

**Reformulation**

Dualize \( \ell \) with its Fenchel conjugate \( \ell^* \) and solve for \( \alpha \)

\[
\begin{align*}
\min_{\alpha \in \mathbb{R}^+} & \quad \sup_{\theta \in \mathbb{R}^+} \left\{ \psi(\theta) - \theta \ell^*(\theta) - \frac{1}{2} \psi'(\theta) \langle K \odot \theta \rangle \right\}
\end{align*}
\]

**Rounding**

Simply use \( \eta \) (ignoring \( M \)) and re-solve for \( \alpha \).

**Reoptimization (optional)**

Local improvement by alternating between \( \eta \) and \( \alpha \).

### Idea: Variational M-estimators

**Variational loss**

\[
\rho(\hat{r}) = \begin{cases} 
\psi(\hat{r}) & \text{if } \hat{r} = 0 \\
\theta \psi(\hat{r}) + \psi'(\hat{r}) \hat{r} & \text{if } \hat{r} \neq 0
\end{cases}
\]

where \( r = y - f(x) \) is the residual, \( \eta \) is the outlier indicator, and \( \psi(\eta) \) is a penalty for outlier attribution.

**Variational M-estimator**

\[
\min_{\eta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \left[ \eta_i (y_i - f(x_i)) + \psi(\eta_i) \right] + \frac{\lambda}{2} \| \eta \|_1,
\]

and its dual

\[
\min_{\eta \in \mathbb{R}^n} \psi(\eta) + \frac{\lambda}{2} \| \eta \|_1 + \frac{1}{2} \| \eta \|_2^2.
\]

where \( K := \odot (x_i, x_j) \) is the kernel matrix.

Not jointly convex in \( \alpha \) and \( \eta \). Alternating minimization cannot guarantee local optimality.

### A Polynomial-time Form of Robust Regression

Yaoliang Yu, Özlem Aslan and Dale Schuurmans

Department of Computing Science, University of Alberta

\{yaoliang, ozlem, dale\}@cs.ualberta.ca

---

**Summary**

Is consist + robust + tractable regression achievable?

- Properties true or false
- Robustness 1 1 1 0 1 1 1 1
- M-estimators 1 1 1 1 1
- Achievable 0 1 1 1

**But, we prove**

**Theorem 1:** (Regularized) M-estimators based on any (non-constant) convex loss cannot have bounded response if the kernel is unbounded.

- This essentially rules out (fixed) convex losses.

**Theorem 2:** M-estimators based on any bounded (non-convex) loss are \( \text{NP-hard} \) to minimize.

- This essentially rules out (fixed) bounded losses.

Therefore it is unlikely that all desired properties can be achieved by M-estimators.

---

**Convex Relaxation**

Dualize \( \ell \) with its Fenchel conjugate \( \ell^* \) and solve for \( \alpha \)

\[
\min_{\alpha \in \mathbb{R}^+} \sup_{\theta \in \mathbb{R}^+} \left\{ \psi(\theta) - \theta \ell^*(\theta) - \frac{1}{2} \psi'(\theta) \langle K \odot \theta \rangle \right\}
\]

**Consistency?**

**Theorem 4:** Assume \( \ell \) is Lipschitz and \( \psi \) is bounded. Assume that the data is generated from a mixture of inliers and outliers, where \( P(\text{inlier}) \gg P(\text{outlier}) \). Then the convex relaxation of VM is loss consistent.

**Computational tractability?**

The convex relaxation is apparently tractable, as long as \( \psi \) and \( \ell \) are convex (and evaluable in poly-time).