Better generalization with less data using robust gradient descent

Matthew J. Holland ¹  Kazushi Ikeda ²

¹Osaka University
²Nara Institute of Science and Technology
**Distribution robustness**

In practice, the learner does not know what kind of data it will run into in advance.

**Q:** Can we expect to be able to use the same procedure for a wide variety of distributions?
A natural baseline: ERM

Empirical risk minimizer:

\[
\hat{w}_{\text{ERM}} \in \arg \min_w \frac{1}{n} \sum_{i=1}^{n} l(w; z_i) \approx \arg \min_w R(w)
\]

Risk:

\[
R(w) := \int l(w; z) \, d\mu(z)
\]

When data is sub-Gaussian, ERM via (S)GD is “optimal.”

(Lin and Rosasco, 2016)

How does ERM fare under much weaker assumptions?
ERM is not distributionally robust

Consider iid $x_1, \ldots, x_n$ with $\text{var}_\mu x = \sigma^2$.

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

**Ex.** Normally distributed data.

$$|\bar{x} - \mathbf{E} x| \leq \sigma \sqrt{\frac{2 \log(\delta^{-1})}{n}}$$

**Ex.** All we know is $\sigma^2 < \infty$.

$$\frac{\sigma}{\sqrt{n\delta}} \left(1 - \frac{e \delta}{n}\right)^{(n-1)/2} \leq |\bar{x} - \mathbf{E} x| \leq \frac{\sigma}{\sqrt{n\delta}}$$

If unlucky, lower bound holds w/ prob. at least $\delta$.

(Catoni, 2012)
Intuitive approach: construct better feedback

\[ \hat{x}_M := \arg \min_{u \in \mathbb{R}} \sum_{i=1}^{n} \rho \left( \frac{x_i - u}{s} \right) \]

*Figure:* Different choices of \( \rho \) (left) and \( \rho' \) (right): \( \rho(u) \) as \( u^2/2 \) (cyan), as \( |u| \) (green), and as \( \log \cosh(u) \) (purple).
Intuitive approach: construct better feedback

Assuming only that the variance $\sigma^2$ is finite,

$$|\hat{x}_M - E \, x| \leq 2\sqrt{\frac{2 \log(\delta^{-1})}{n}} \sigma$$

at probability $1 - \delta$ or greater.

(Catoni, 2012)

Compare:

$$\bar{x}: \sqrt{\delta^{-1}} \text{ vs. } \hat{x}_M: 2\sqrt{2 \log(\delta^{-1})}$$
Previous work considers robustified objectives

\[
L_M(w) := \arg \min_{u \in \mathbb{R}} \sum_{i=1}^{n} \rho \left( \frac{l(w; z_i) - u}{s} \right)
\]

\[
\downarrow
\]

\[
\widehat{w}_{BJL} = \arg \min_{w} L_M(w).
\]

(Brownlees et al., 2015)

+ General purpose distribution-robust risk bounds.
+ Can adapt to a “guess and check” strategy.

– Defined implicitly, difficult to optimize directly.
– Most ML algorithms only use first-order information.

(Holland and Ikeda, 2017b)
Our approach: aim for risk gradient directly

Early work by Holland and Ikeda (2017a) and Chen et al. (2017).
Later evolutions in Prasad et al. (2018); Lecué et al. (2018).
Our approach: aim for risk gradient directly

Early work by Holland and Ikeda (2017a) and Chen et al. (2017).
Later evolutions in Prasad et al. (2018); Lecué et al. (2018).
Our approach: aim for risk gradient directly

Early work by Holland and Ikeda (2017a) and Chen et al. (2017).
Later evolutions in Prasad et al. (2018); Lecué et al. (2018).
Our proposed robust GD

Key sub-routine:

\[
\hat{g}(w) = (\hat{\theta}_1(w), \ldots, \hat{\theta}_d(w)) \approx \nabla R(w)
\]

\[
\hat{\theta}_j := \arg \min_{\theta \in \mathbb{R}^n} \sum_{i=1}^{n} \rho \left( \frac{l_j'(w; z_i) - \theta}{s_j} \right), \quad j \in [d].
\]

Plug into descent update:

\[
\hat{w}_{(t+1)} = \hat{w}_{(t)} - \alpha_{(t)} \hat{g}(\hat{w}_{(t)}).
\]

Variance-based scaling:

\[
s_j^2 = \text{var} l_j'(w; z)n \frac{n \log (2 \delta^{-1})}{\log (2 \delta^{-1})}.
\]
Our proposed robust GD

+ Guarantees requiring only finite variance:

\[ O \left( \frac{d \left( \log(d\delta^{-1}) + d \log(n) \right)}{n} \right) + O \left( (1 - \alpha)^T \right) \]

+ Theory holds as-is for implementable procedure.

+ Small overhead; fixed-point sub-routine converges quickly.

  – Naive coordinate-wise strategy leads to sub-optimal guarantees; in principle, can do much better.

  (Lugosi and Mendelson, 2017, 2018)

  – If non-convex, useful exploration may be constrained.
Looking ahead

Q: Can we expect to be able to use the same procedure for a wide variety of distributions?
Q: Can we expect to be able to use the same procedure for a wide variety of distributions?

A: Yes, using robust GD. However, it is still far from optimal.

Catoni and Giulini (2017); Lecué et al. (2018); Minsker (2018)

Can we get nearly sub-Gaussian estimates in linear time?


