Better generalization with less data using robust gradient descent

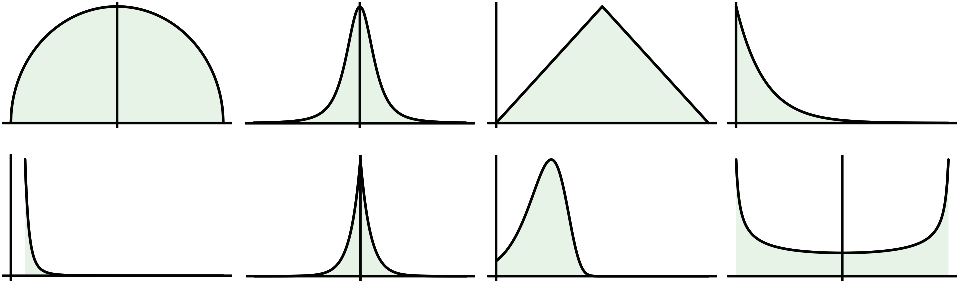
Matthew J. Holland [1] **Kazushi Ikeda** [2]

[1] Osaka University

[2] Nara 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:

$$\widehat{\boldsymbol{w}}_{\text{ERM}} \in \operatorname*{arg\,min}_{\boldsymbol{w}} \frac{1}{n} \sum_{i=1}^{n} l(\boldsymbol{w}; \boldsymbol{z}_i)$$
$$\approx \operatorname*{arg\,min}_{\boldsymbol{w}} R(\boldsymbol{w})$$

Risk:

$$R(\boldsymbol{w}) := \int l(\boldsymbol{w}; \boldsymbol{z}) \, d\mu(\boldsymbol{z})$$

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 $\operatorname{var}_\mu x = \sigma^2$.

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

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 δ.

(Catoni, 2012)

Intuitive approach: construct better feedback

$$\widehat{x}_{\mathsf{M}} := \arg\min_{u \in \mathbb{R}} \sum_{i=1}^{n} \rho\left(\frac{x_i - u}{s}\right)$$

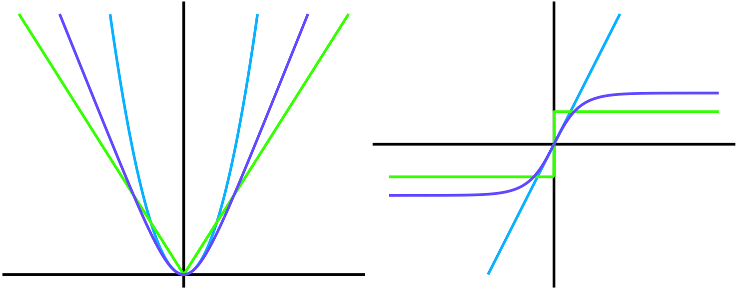


Figure: Different choices of ρ (left) and ρ' (right): $\rho(u)$ as $u^2/2$ (cyan), as $|u|$ (green), and as $\log\cosh(u)$ (purple).

Intuitive approach: construct better feedback

Assuming only that the variance σ^2 is finite,

$$|\widehat{x}_{\mathsf{M}} - \mathbf{E}\, x| \leq 2\sqrt{\frac{2\log(\delta^{-1})}{n}}\sigma$$

at probability $1-\delta$ or greater.

(Catoni, 2012)

Compare:

$$\bar{x}\text{: } \sqrt{\delta^{-1}} \quad \text{vs.} \quad \widehat{x}_{\mathsf{M}}\text{: } 2\sqrt{2\log(\delta^{-1})}$$

Previous work considers robustified objectives

$$L_{\mathsf{M}}(\boldsymbol{w}) := \operatorname*{arg\,min}_{u \in \mathbb{R}} \sum_{i=1}^{n} \rho\left(\frac{l(\boldsymbol{w}; \boldsymbol{z}_i) - u}{s}\right)$$

$$\downarrow$$

$$\widehat{\boldsymbol{w}}_{\mathsf{BJL}} = \operatorname*{arg\,min}_{\boldsymbol{w}} \, L_{\mathsf{M}}(\boldsymbol{w}).$$

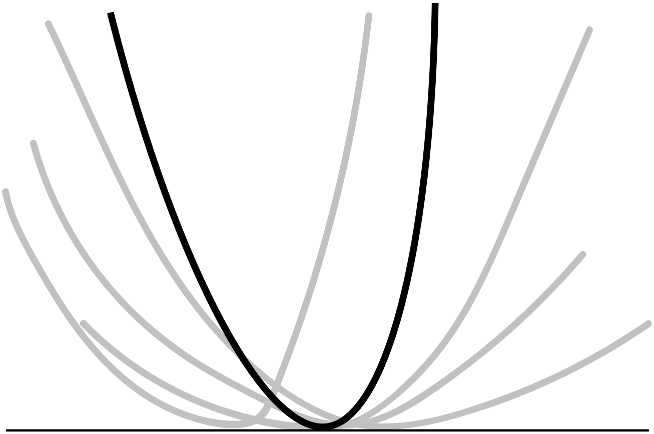
(Brownlees et al., 2015)

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

(Holland and Ikeda, 2017b)

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

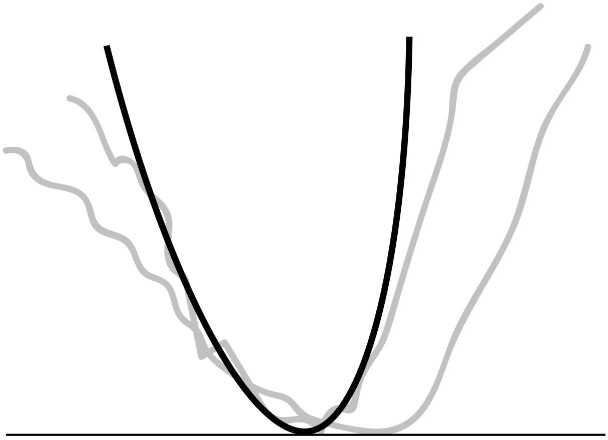
Our approach: aim for risk gradient directly



Early work by Holland and Ikeda (2017a) and Chen et al. (2017).

Later evolutions in Prasad et al. (2018); Lecué et al. (2018).

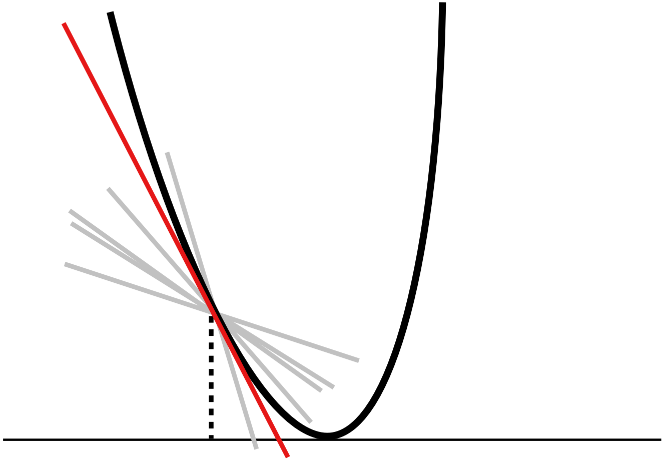
Our approach: aim for risk gradient directly



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:

$$\widehat{\boldsymbol{g}}(\boldsymbol{w}) = \left(\widehat{\theta}_1(\boldsymbol{w}), \ldots, \widehat{\theta}_d(\boldsymbol{w})\right) \approx \nabla R(\boldsymbol{w})$$

$$\widehat{\theta}_j := \operatorname*{arg\,min}_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \rho\left(\frac{l'_j(\boldsymbol{w}; \boldsymbol{z}_i) - \theta}{s_j}\right), \quad j \in [d].$$

Plug into descent update:

$$\widehat{\boldsymbol{w}}_{(t+1)} = \widehat{\boldsymbol{w}}_{(t)} - \alpha_{(t)} \, \widehat{\boldsymbol{g}}(\widehat{\boldsymbol{w}}_{(t)}).$$

Variance-based scaling:

$$s_j^2 = \frac{\operatorname{var} l'_j(\boldsymbol{w}; \boldsymbol{z}) n}{\log\left(2\delta^{-1}\right)}.$$

Our proposed robust GD

+ Guarantees requiring only finite variance:

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

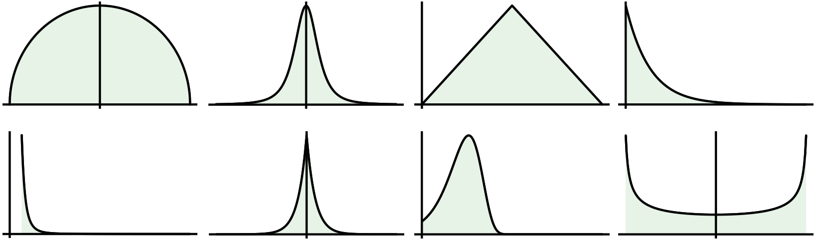
+ Theory holds as-is for implementable procedure.
+ Small overhead; fixed-point sub-routine converges quickly.

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

(Lugosi and Mendelson, 2017, 2018)

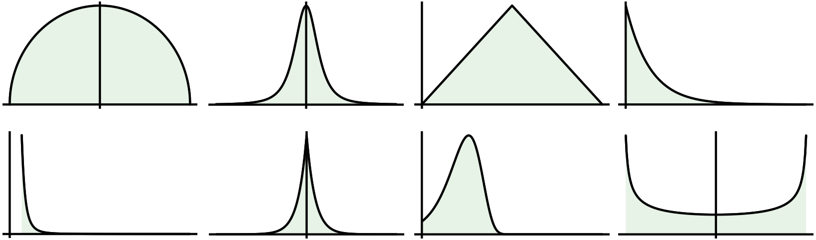
– If non-convex, useful exploration may be constrained.

Looking ahead



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

Looking ahead



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?

References

Brownlees, C., Joly, E., and Lugosi, G. (2015). Empirical risk minimization for heavy-tailed losses. *Annals of Statistics*, 43(6):2507–2536.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185.

Catoni, O. and Giulini, I. (2017). Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression. *arXiv preprint arXiv:1712.02747*.

Chen, Y., Su, L., and Xu, J. (2017). Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. *arXiv preprint arXiv:1705.05491*.

Holland, M. J. and Ikeda, K. (2017a). Efficient learning with robust gradient descent. *arXiv preprint arXiv:1706.00182*.

Holland, M. J. and Ikeda, K. (2017b). Robust regression using biased objectives. *Machine Learning*, 106(9):1643–1679.

Lecué, G., Lerasle, M., and Mathieu, T. (2018). Robust classification via MOM minimization. *arXiv preprint arXiv:1808.03106*.

Lin, J. and Rosasco, L. (2016). Optimal learning for multi-pass stochastic gradient methods. In *Advances in Neural Information Processing Systems 29*, pages 4556–4564.

Lugosi, G. and Mendelson, S. (2017). Sub-gaussian estimators of the mean of a random vector. *arXiv preprint arXiv:1702.00482*.

References (cont.)

Lugosi, G. and Mendelson, S. (2018). Near-optimal mean estimators with respect to general norms. *arXiv preprint arXiv:1806.06233*.

Minsker, S. (2018). Uniform bounds for robust mean estimators. *arXiv preprint arXiv:1812.03523*.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2018). Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*.