Learning bounds for Risk-sensitive Learning

Jaeho Lee, Sejun Park, Jinwoo Shin Korea Advanced Insitute of Science and Technology (KAIST)

Mail: jaeho-lee@kaist.ac.kr
Github: jaeho-lee/oce
Twitter: @jaeho_lee_

arXiv:2006.08138

TL;DR. We formulate risk-averse/seeking learning algorithms as an empirical **OCE** minimization, and give theoretical generalization guarantees.

Motivation. Robust/Fair ML algorithms

Robust/Fair ML algorithms discriminate samples, based on their losses.

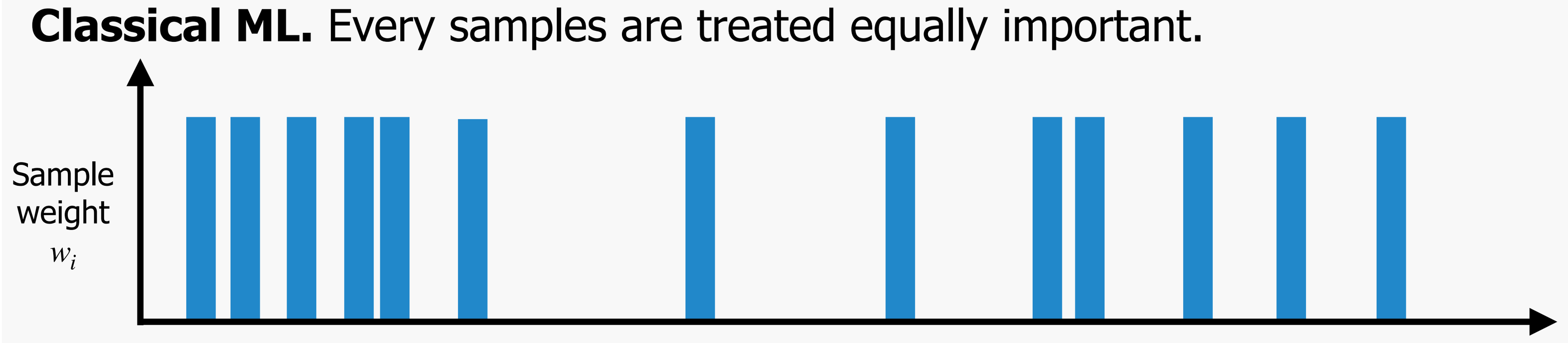
Fact. ML algorithms can be viewed as "minimizing the weighted sum" of losses: Given the training data $Z_1, Z_2, \ldots, Z_n$, we find a parameter $\theta \in \Theta$ (e.g. neural network weights) that achieve

$$\min_{\theta \in \Theta} \sum_{i=1}^{n} w_i \cdot l(Z_i; \theta)$$

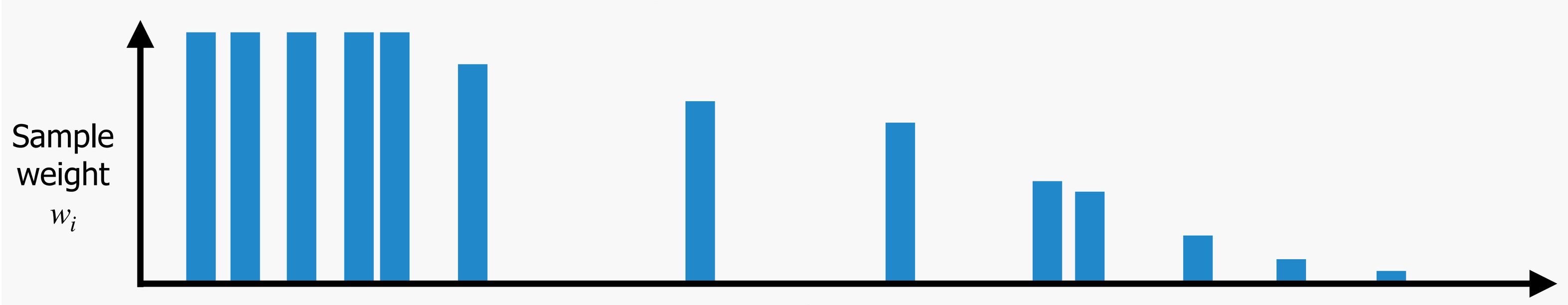
for some weights $w_1, w_2, \ldots, w_n$.

The weights are typically…

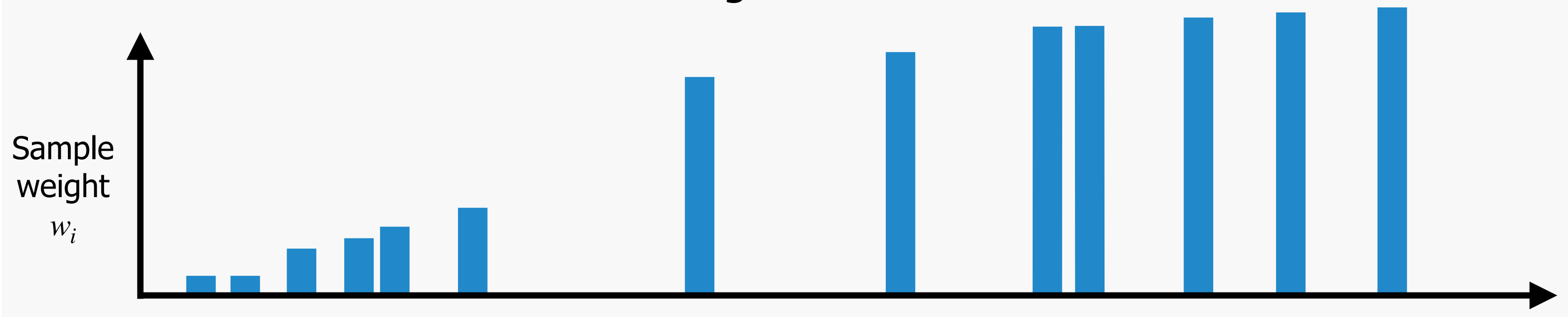
Classical ML. Every samples are treated equally important.



Robust ML. High-loss samples are viewed as "outliers," and are disregarded or considered less important.



Fair ML. Reducing the loss of high-loss samples is prioritized, to mitigate the sense of unfairness among individuals.



What we know. The statistical learning theory of Vapnik & Chervonenkis handles the first ("uniform weight") scenario very well; the behavior of mean is very well understood!

Question. How do we handle other two?

The theoretical framework for analyzing robust / fair ML algorithms.

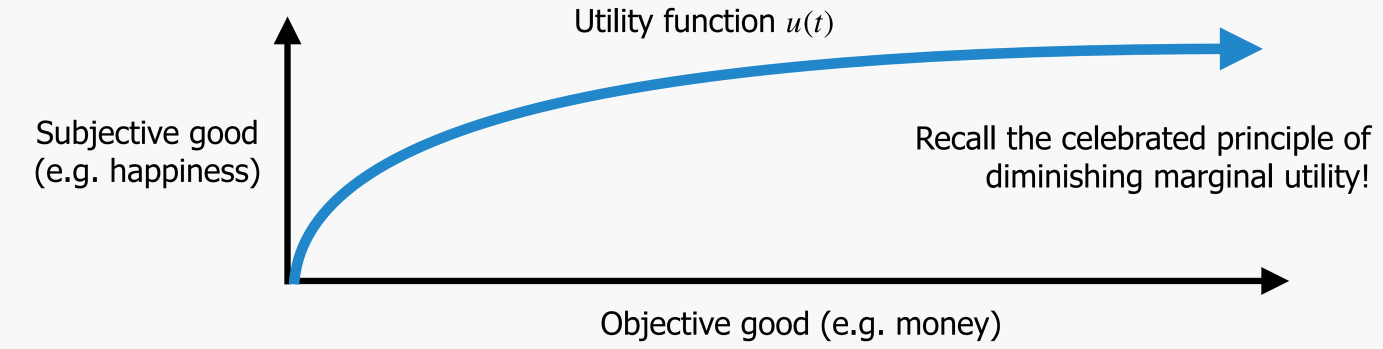
Our answer. We can extend the economic/financial theory on "risk-sensitivity" to handle the other scenarios.

In particular, we find the notion of **OCE** very useful! (optimized certainty equivalents)

Background. OCE… what's that?

A brief prehistory of Optimized Certainty Equivalents (OCE)

1947. The idea behind the OCE originates from the famous "utility theory" of von Neumann and Morgenstern, where the irrationality of an individual decision-maker is explained in terms of its innate utility function.



1986. Ben-Tal and Teboulle formalizes the notion of optimized certainty equivalent (OCE), as a measure of risk (i.e., loss aggregation method) that accounts for risk-averse/seeking tendency of the decision-maker:

$$\mathrm{oce}(\theta, P) \triangleq \inf_{\lambda \in \mathbb{R}} \left\{ \lambda + \mathbf{E}_P \left[\phi(l(Z;\theta) - \lambda) \right] \right\}$$

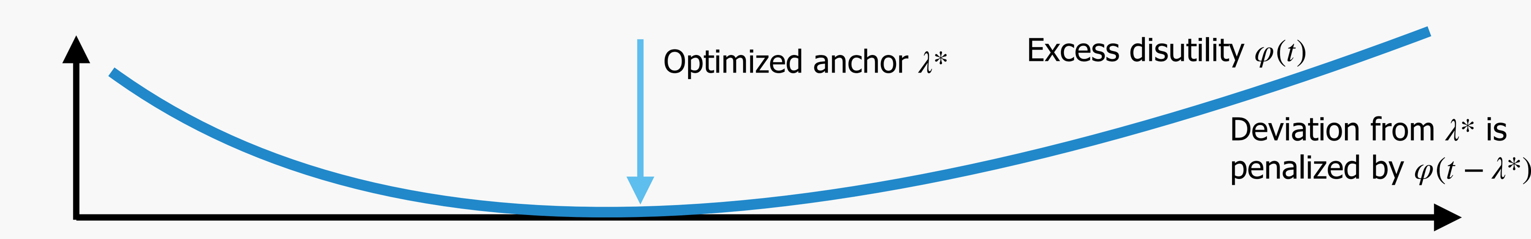
for some disutility function $\phi(t) = -u(t)$.

In other words, OCE is the maximized "present + future value," where the future value is uncertain and follows the utility curve.

2020. Lee and friends (us!) realize that OCE can be also written as

$$\mathrm{oce}(\theta, P) = \mathbf{E}_P[l(Z;\theta)] + \inf_{\lambda \in \mathbb{R}} \mathbf{E}_P \left[\varphi(l(Z;\theta) - \lambda) \right]$$

for the "bowl-shaped" excess disutility function $\varphi(t) = \phi(t) - t$.



That is, OCE is "mean loss + deviation penalty," representing many fair-ML -like algorithms implicitly minimizing CVaR, entropic risk, or variance penalty.

Result#1. Inverting OCE for Robust ML

We newly define inverted OCE to formally address robust ML algorithms

Inverse. We propose the "inverted OCE" to characterize the robust-ML-like algorithms which disregard high-loss samples.

$$\overline{\mathrm{oce}}(\theta, P) \triangleq \sup_{\lambda \in \mathbb{R}} \left\{ \lambda + \mathbf{E}_P \left[\phi(\lambda - l(Z;\theta)) \right] \right\}$$

Prop. 1. (Informal) As a special case, this modification incorporates algorithms that ignore high-loss samples.

Prop. 2. (Informal) Inverted OCE risks are more robust than the average loss, in terms of the influence function.

Result#2. Performance guarantees

Rademacher complexity bounds for empirical OCE minimizers.

EOM. Similar to ERM (empirical risk minimization), we consider EOM procedure:

$$\hat{\theta}_{\mathrm{eom}} \triangleq \mathrm{argmin}_\theta\, \mathrm{oce}(\theta, P_n)$$

where P_n is the empirical distribution of the training dataset.

Thm. 1. (Informal) We have the excess OCE risk bound: with high probability,

$$\mathrm{oce}(\hat{\theta}_{\mathrm{eom}}, P) - \inf_\theta \mathrm{oce}(\theta, P) = \mathcal{O}\left(\frac{\mathrm{Lip}(\phi) \cdot \mathrm{comp}(\Theta)}{\sqrt{n}} \right)$$

where $\mathrm{comp}(\Theta)$ denotes the Rademacher complexity of Θ.

Thm. 2. (Informal) We have the excess mean loss bound: with high probability,

$$\mathbf{E}l(\hat{\theta}_{\mathrm{eom}}, P) - \inf_\theta \mathbf{E}l(\theta, P) = \mathcal{O}\left(\frac{\mathrm{comp}(\Theta)}{\sqrt{n}} \right) + \varepsilon$$

where ε is a small term proportional to the loss standard deviation of the optimal hypothesis.

Note. We also prove the similar results for empirical inverted OCE minimizers!

Result#3. Algorithmic implications

Sample variance penalization can be used for OCE minimization, provably.

Remark. In proving Thm. 2., we observe that OCE minimization is almost equivalent to the sample variance penalization procedure.

Batch-based sample variance penalization is often less noisier than the batch-based OCE minimization (using full-batch information).

Idea. Why don't we use sample variance penalization as a baseline method for the empirical OCE minimization?

Result. The proposed baseline outperforms naïve batch CVaR minimization!