Towards Reciprocal Learning Theory

Generalizing From Self-Selected Samples

Julian Rodemann[1], Christoph Jansen[2], Georg Schollmeyer[1], James Bailie[3]

2nd Workshop on Learning Under Weakly Structured Information

Tübingen

[1]Department of Statistics, LMU Munich
[2]School of Computing & Communications, Lancaster University
[3]Department of Statistics, Harvard University

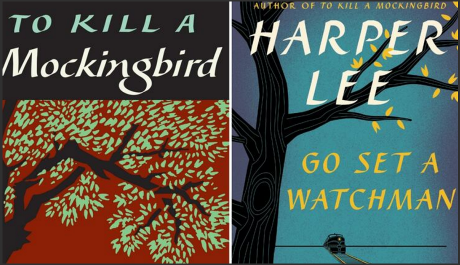


Image Credits: CBS, Prime Video

Contents

Contents

Motivating Example: Self-Training[1]

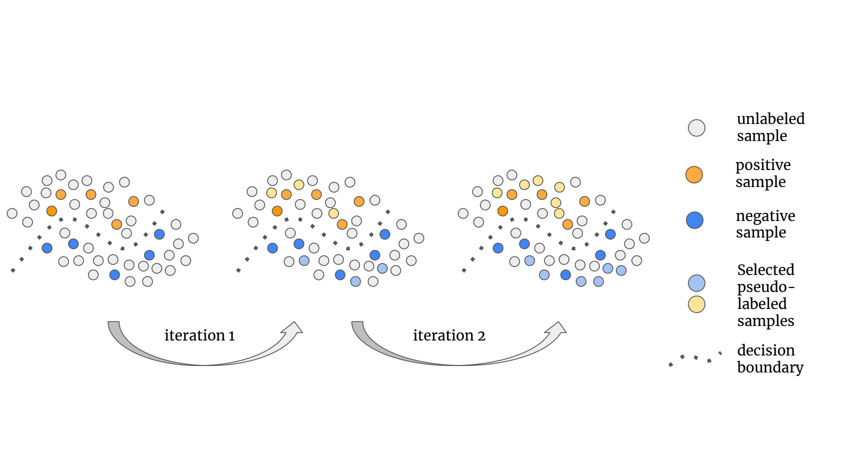


Figure 1: Sketch of Self-Training for Binary Classification

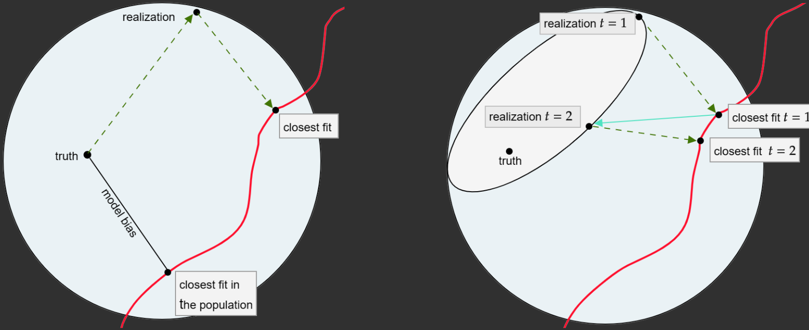
[1] Other names: Pseudo-Labeling, Self-Labeling.

Motivation

- Growing interest in experimental design and subsampling[2]
- We demonstrate that a wide range of (online) learning algorithms *implicitly* design experiments
- *reciprocal* relationship between data and parameters: These algorithms not only learn parameters from data, but also vice versa
  - self-training (semi-supervised learning)
  - active learning
  - boosting
  - multi-armed bandits
  - superset learning
  - Bayesian optimization
  - ...

[2] (Lang, Vijayaraghavan, and Sontag 2022; Liu et al. 2023; Malte Nalenz, Rodemann, and Thomas Augustin 2024; Pooladzandi, Davini, and Mirzasoleiman 2022; Rodemann 2024; Rodemann, Fischer, et al. 2022; Stolz 2023; Yin et al. 2024)

A Visual Perspective



(a) Classical statistical learning

(b) Reciprocal Learning

Figure 2: (a) Classical statistical learning fits a model from the model space (restricted by red curve) to a realized sample from the sample space (blue-grey), see Hastie, Tibshirani, and Friedman 2009, Figure 7.2. (b) In reciprocal learning, the sample changes *in response to* the model fit.

What's a fit?

Definition (Risk)

Let $\theta \in \Theta$ a parameter vector. The **risk** of θ on $Z = (X, Y) \sim P$ is

$$\mathscr{R}(P, \theta) := \mathbb{E}_P \left[\ell(Y, X, \theta)\right] = \int_{\mathcal{Z}} \ell(y, x, \theta) \mathrm{d}P(z),$$

where $\ell : \mathcal{Y} \times \mathcal{X} \times \Theta \to \mathbb{R}$ is a loss function and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

Definition (Empirical Risk Minimizer)

Denote by $(x_1, y_1), \ldots, (x_n, y_n)$ a sample with empirical law $\hat{\mathbb{P}}_0$. Call

$$\hat{\theta}_t \in \underset{\theta \in \Theta}{\arg\min}\, \mathscr{R}\left(\hat{\mathbb{P}}_0, \theta\right) = \underset{\theta}{\arg\min}\, \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i, \theta)$$

the **empirical risk minimizer** in iteration $t \in \{1, \ldots, T\}$.

Reciprocal Learning

Call an algorithm *reciprocal* if ...

1. ... it performs empirical risk minimization (ERM) ...
2. ... iteratively ...
3. ... on a sample that depends ...
   - ... on the previous ERM solution ...
   - ... and the previous sample.

Reciprocal Learning

Definition (Sample Adaptation)

Denote by Θ a parameter space and by $\mathscr{P}$ a space of (joint) probability distributions of X and Y. Call

$$f_s : \begin{cases} \Theta \times \mathscr{P} & \longrightarrow \mathscr{P} \\ (\hat{\theta}_t, \hat{\mathbb{P}}_t(Y, X)) & \longmapsto \hat{\mathbb{P}}_{t+1}(Y, X) \end{cases}$$

the sample adaptation function.

simple example

Reciprocal Learning

Definition (Reciprocal Learning)

With f_s, Θ, $\mathscr{P}$, X, and Y as above, we define

$$R : \begin{cases} \Theta \times \mathscr{P} & \longrightarrow \Theta \times \mathscr{P}; \\ (\hat{\theta}_t, \hat{\mathbb{P}}_t) & \longmapsto (\hat{\theta}_{t+1}, \hat{\mathbb{P}}_{t+1}) \end{cases}$$

as reciprocal learning, where

$$\hat{\mathbb{P}}_{t+1} = f_s(\hat{\theta}_t, \hat{\mathbb{P}}_t)$$

and

$$\hat{\theta}_{t+1} = \operatorname*{arg\,min}_{\theta \in \Theta} \mathscr{R}(\hat{\mathbb{P}}_{t+1}, \theta)$$

Contents

Yes, it can!

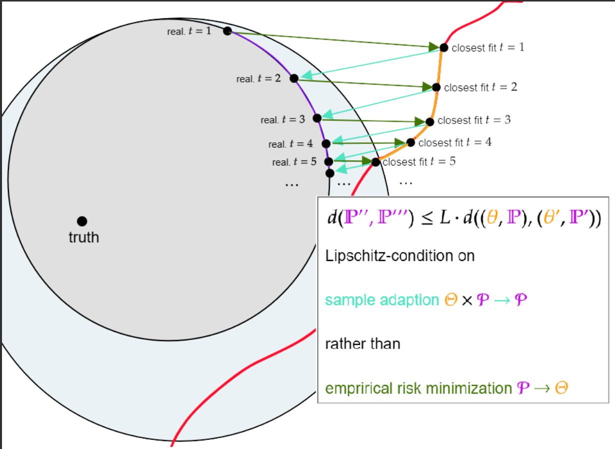


Figure 3: Idea: Bound the change in the data (purple) by the change in the model (yellow).

Convergence of Reciprocal Learning

Theorem (Convergence of Reciprocal Learning, Informal)

If the sample adaptation f_s is sufficiently Lipschitz (w.r.t. Wasserstein-2 on $\mathcal{P}$ and L2 on Θ) and the loss is strongly convex and continuously differentiable, the iterates $R_t = (\theta_t, \mathbb{P}_t)$ of reciprocal learning R converge to the limit $(\theta_c, \mathbb{P}_c)$ point-wise at a linear rate.

Proof idea:

- Show that $R : \Theta \times \mathcal{P} \to \Theta \times \mathcal{P}$ is a contraction
- Convergence is a fixed-point condition on R
- Apply Banach fixed-point theorem

Sufficient Conditions for Lipschitz-Continuity of f_s

- When is the sample adapation sufficiently Lipschitz?
    - continuous predictions of y
    - regularized or randomized selection of x

Regularized Data Selection

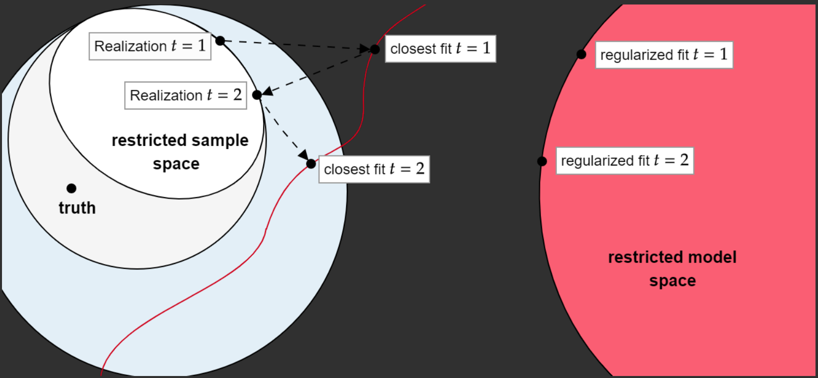


Figure 4: Data regularization is symmetrical to classical model regularization, see illustration in “The Elements of Statistical Learning” Hastie, Tibshirani, and Friedman 2009, Figure 7.2.

Some Corollaries

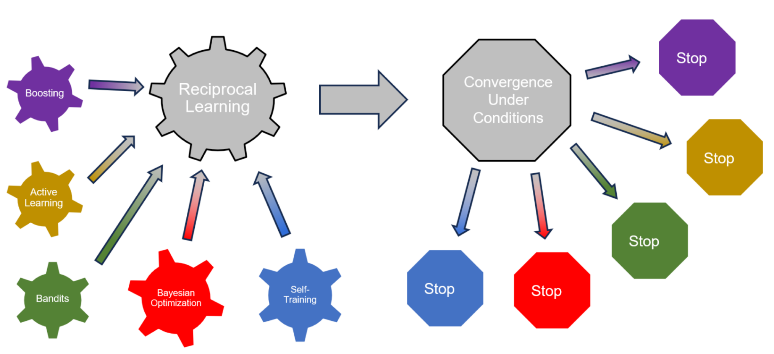


Figure 5: Reciprocal Learning generalizes (left) machine learning algorithms. Corollaries (right) of convergence results give rise to theory-informed bandits, active learning, self-training etc. algorithms that shall converge.

References: Reciprocal Learning

Paper:

Summary:

Julian Rodemann, Christoph Jansen, Georg Schollmeyer (2024). Reciprocal Learning, **First Workshop on Learning Under Weakly Structured Information (LUWSI)**, Munich.

Julian Rodemann, Christoph Jansen, Georg Schollmeyer (2024). Reciprocal Learning, **Advances in Neural Information Processing Systems (NeurIPS)**, Vancouver.

Contents

Can Reciprocal Learning Generalize?

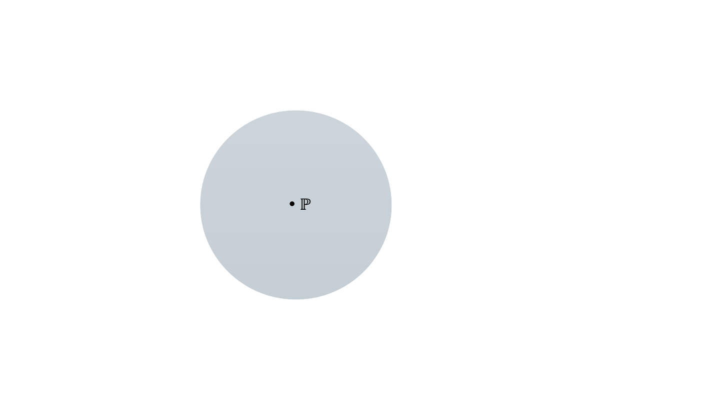


Figure 6: Sample Space

Can Reciprocal Learning Generalize?

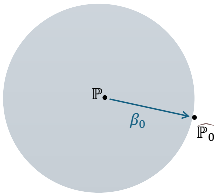


Figure 7: Bound on Wasserstein distance between law $\mathbb{P}$ and *i.i.d.* sample $\hat{\mathbb{P}}_0$: $W_p(\mathbb{P}, \hat{\mathbb{P}}_0) \leq \beta_0$ (Fournier and Guillin 2015)

Can Reciprocal Learning Generalize?

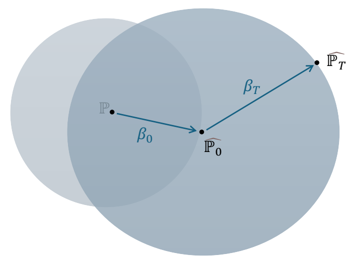


Figure 8: Reciprocal distortion bound $W_p(\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_T) \leq \beta_T$

Can Reciprocal Learning Generalize?

Lemma (Reciprocal Distortion Bound)

Let $\hat{\mathbb{P}}_T$ be the empirical distribution of training data at iteration T in reciprocal learning and $\hat{\mathbb{P}}_0$ the initial one. Denote by $W_p(\cdot,\cdot)$ the p-Wasserstein distance and by $\sup_{z,z'} d_{\mathcal{Z}}(z,z')^p < \infty$ the diameter bound. It holds

$$W_p(\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_T) \leq \frac{L_s^T - 1}{L_s - 1} \frac{\sup_{z,z'} d_{\mathcal{Z}}(z,z')^p}{n} := \beta_T,$$

where $0 < L_s < 1$ is the Lipschitz constant of the sample adaptation function f_s.

Observe $\beta_\infty = \frac{\sup_{z,z'} d_{\mathcal{Z}}(z,z')^p}{(1-L_s)n}$.

Can Reciprocal Learning Generalize?

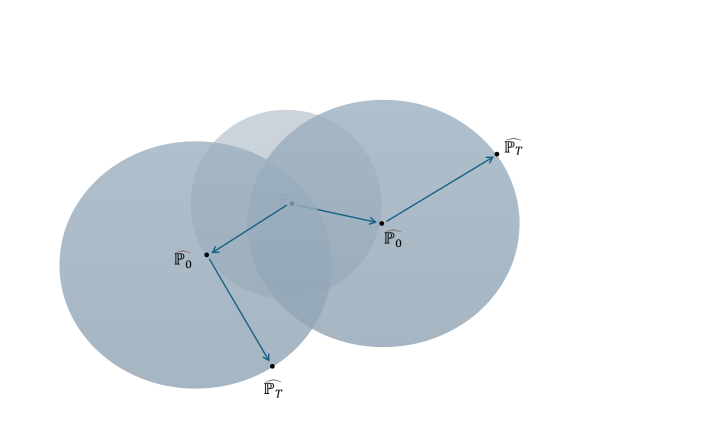


Figure 9: Wasserstein Ball for Reciprocal Learning

Can Reciprocal Learning Generalize?

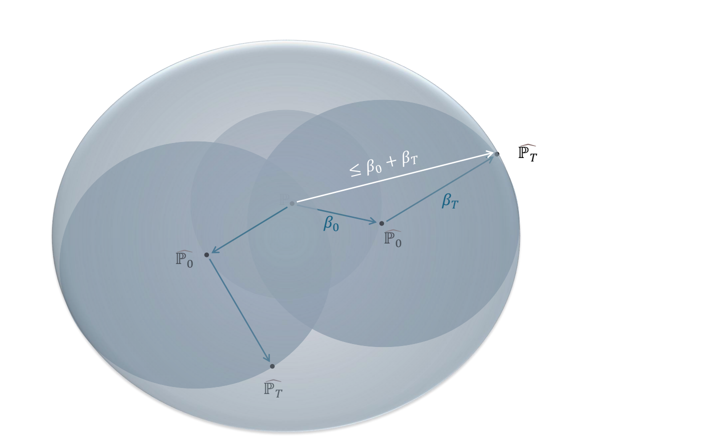


Figure 10: Wasserstein Ball for Reciprocal Learning

Generalization Bounds for Reciprocal Learning

Theorem (Generalization Gap)

Assume Θ is compact. Let $\hat{\theta}_c$ be the solution of reciprocal learning with corresponding $\hat{\mathbb{P}}_c$. It holds

$$\mathscr{R}(\mathbb{P}, \hat{\theta}_c) \leq \mathscr{R}(\hat{\mathbb{P}}_c, \hat{\theta}_c) + L_\ell \underbrace{\left(\frac{\log\left(4C_a/\delta\right)}{C_b n} \right)^{p/d}}_{\beta_0} + L_\ell \underbrace{\frac{\sup_{z,z'} d_{\mathcal{Z}}(z, z')^p}{(1 - L_s)n}}_{\beta_\infty}$$

with probability of at least $1 - \delta$ and Lipschitz-constant L_ℓ of the loss. Constants C_a, C_b and dimension d are properties of $\mathcal{Z}$.

Generalization Bounds for Reciprocal Learning

Theorem (Excess Risk Bound, Informal)

Assume Θ is compact. The excess risk $\mathscr{R}(\mathbb{P}, \hat{\theta}_c) - \inf_{\theta \in \Theta} \mathscr{R}(\mathbb{P}, \theta)$ of reciprocal learning is upper bounded with high probabiliy

Proof Idea:

- Bound $\mathscr{R}(\mathbb{P}, \hat{\theta}_c) - \mathscr{R}(\hat{\mathbb{P}}_0, \hat{\theta}_c)$ via Kantorovich-Rubinstein lemma
- Tricky part: bound $\mathscr{R}(\hat{\mathbb{P}}_0, \hat{\theta}_c) - \mathscr{R}(\hat{\mathbb{P}}_0, \hat{\theta}_0)$
- Bound $\mathscr{R}(\hat{\mathbb{P}}_0, \hat{\theta}_0) - \inf_{\theta \in \Theta} \mathscr{R}(\mathbb{P}, \theta)$ via a standard symmetrization argument

Generalization Bounds for Reciprocal Learning

Theorem (Excess Risk Bound)

Assume Θ is compact. The excess risk $\mathscr{R}(\mathbb{P}, \hat{\theta}_c) - \mathscr{R}_\Theta$ of reciprocal learning is upper bounded by

$$L_\ell \left(\frac{\log(4C_a/\delta)}{C_b n} \right)^{p/d} + \frac{(1 + L_a) L_\ell \sup_{z,z'} d_{\mathcal{Z}}(z, z')^p}{n(1 - L_s \max\{1, \frac{\kappa}{\gamma}\})} + \frac{1}{\sqrt{n}} \left(24\mathfrak{C}(\mathcal{F}) + \sqrt{2\ln(1/\delta)} \right),$$

with probability greater or equal than $1 - \frac{\delta}{2}$, where C_a, C_b are again constants depending on p, d, and $\sup_{z,z'} d_{\mathcal{Z}}(z, z')$, L_ℓ is the Lipschitz-constant of the loss, κ and γ are from conditions 2 and 3. L_a shall denote the Lipschitz-constant of $\mathcal{P} \to \Theta : P \mapsto \arg\min_{\theta \in \Theta} \mathscr{R}(P, \theta)$. $\mathfrak{C}_{L_2}(\mathcal{F})$ is the covering entropy integral of a hypothesis class $\mathcal{F} := \{f_\theta : \mathcal{X} \to \mathcal{Y} \mid \theta \in \Theta\}$ with $f_\theta \neq f_{\theta'}$ for $\theta \neq \theta'$.

LMU

Anytime Valid Bounds

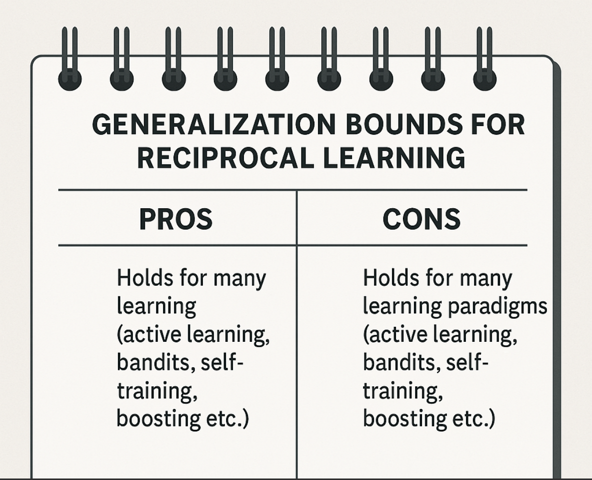
- What if we stop earlier?
- Good news: Bounds still hold for compact Θ if data selection is regularized (strong convexity of loss not required)

Contents

Summary

- Reciprocal learning fruitfully generalizes active learning, self-training, multi-armed bandits, Bayesian optimization, superset learning, etc.
- Reciprocal learning converges, if change in sample is bounded by the change in model (Lipschitz)
- This can be achieved through *data regularization*
- Generalization requires “good” initial sample and “smooth”/“very continuous” sample adaptation
- Anytime valid bounds allow for stopping criteria with generalization guarantee

Summary


GENERALIZATION BOUNDS FOR RECIPROCAL LEARNING
PROS
CONS
Holds for many learning (active learning, bandits, self-training, boosting etc.)
Holds for many learning paradigms (active learning, bandits, self-training, boosting etc.)

Appendix: Sample Adaptation – Simple Example

Assume $\mathcal{Y} = \{0, 1\}$ and only one data point being added to a sample of size n.

$$\hat{\mathbb{P}}_{t+1}(Y = 1, X = x) = \frac{1(x = ש(\theta)) \cdot צ(ש(\theta), \theta) \; + \; n \, \hat{\mathbb{P}}_t(Y = 1, \; X = x)}{n + 1} \; ,$$

where $צ : \mathcal{X} \times \Theta \longrightarrow \{0, 1\}$ is any function that assigns a label y, potentially based on the model θ, to selected x. Let $ש$ be any function that selects features x given a model θ.

Reciprocal Learning: Examples

- label assignment $\text{צ} : \mathcal{X} \times \Theta \longrightarrow \{0,1\}$
  - Active Learning: Query function $q_y : \mathcal{X} \longrightarrow \{0,1\}$
  - Self-Training (SSL): prediction function $\hat{y} : \mathcal{X} \times \Theta \longrightarrow \{0,1\}$
  - Bandits: $R : \mathcal{A} \longrightarrow \{0,1\}$ reward function
- criterion $c : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}$
  - Active Learning: acquisition function
  - Self-Training (SSL): confidence measures
  - Bandits: policy function[3]
- selection $\text{ש} : \Theta \longrightarrow \mathcal{X}$
  - Active Learning: $\arg\max_x$
  - Self-Training (SSL): $\arg\max_x$
  - Bandits: Thompson sampling, epsilon-greedy (stochastic); upper confidence bound (deterministic)

back

[3] It is often directly defined in terms of action selection probabilities.

Some More Examples

- self-training (semi-supervised learning) (Arazo et al. 2020; Dietrich, Rodemann, and Christoph Jansen 2024; Lee et al. 2013; Li et al. 2020; Rizve et al. 2020; Rodemann, Goschenhofer, et al. 2023; Rodemann, Christoph Jansen, et al. 2023)
- active learning
- multi-armed bandits
- Bayesian optimization (Rodemann 2023; Rodemann and Augustin 2024; Rodemann and Thomas Augustin 2021, 2022; Rodemann, Croppi, et al. 2024)
- superset learning (Hüllermeier and Cheng 2015; Rodemann, Kreiss, et al. 2022)
- ...

Reciprocal Learning Through a Decision-Theoretic Lense

Think of reciprocal learning as sequential decision-making[4]:

$t = 1$ θ_1 solves decision problem $(\Theta, \mathbb{A}_\Theta, \ell)$

a_1 solves decision problem $(\Theta, \mathbb{A}_\mathcal{X}, \ell_{\theta_1})$

$t = 2$ θ_2 solves decision problem $(\Theta, \mathbb{A}_\Theta, \ell_{a_1(\theta_1)})$

a_2 solves decision problem $(\Theta, \mathbb{A}_\mathcal{X}, \ell_{\theta_2})$

$t = 3$ ...

with Θ the set of states of nature (parameter space) as well as action spaces $\mathbb{A}_\Theta = \Theta$ and $\mathbb{A}_\mathcal{X} = \mathcal{X}$ (feature space) for parameter and data selection, respectively.

[4] Notably, addressing machine learning problems from a decision-theoretic point of view has received considerable interest recently C. Jansen, M. Nalenz, et al. 2023; C. Jansen, Schollmeyer, and T. Augustin 2018, 2023; C. Jansen, Schollmeyer, H. Blocher, et al. 2023; Christoph Jansen et al. 2024; Rodemann and Hannah Blocher 2024

Convergence Won't Come For Free

Assumption (Continuous Differentiability in Features)

A loss function $\ell(Y, X, \theta)$ *is said to be continuously differentiable with respect to features if the gradient* $\nabla_X \ell(Y, X, \theta)$ *exists and is* α*-Lipschitz continuous in* θ, x, *and* y *with respect to the L2-norm on domain and codomain.*

Assumption (Continuous Differentiability in Parameters)

A loss function $\ell(Y, X, \theta)$ *is continuously differentiable with respect to parameters if the gradient* $\nabla_\theta \ell(Y, X, \theta)$ *exists and is* β*-Lipschitz continuous in* θ, x, *and* y *with respect to the L2-norm on domain and codomain.*

Convergence Won't Come For Free

Assumption (Strong Convexity)

A loss function $\ell(Y, X, \theta)$ is said to be γ-strongly convex if

$$\ell(y, x, \theta) \geqslant \ell\left(y, x, \theta'\right) + \nabla_\theta \ell\left(y, x, \theta'\right)^\top \left(\theta - \theta'\right) + \frac{\gamma}{2}\left\|\theta - \theta'\right\|_2^2,$$

for all θ, θ', y, x. If $\gamma = 0$, this assumption is equivalent to convexity. If $\ell(y, x, \theta) = \ell(z, \theta)$ with $z \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, then strong convexity is equivalently characterized by the existence of $\gamma > 0$ such that

$$\forall z \in \mathcal{Z} : \nabla_\theta^2 \ell(Z, \theta) \geq \gamma.$$

For examples, see Shalev-Shwartz and Ben-David 2014, Chapter 12.

Conditions for Convergence (ctd.)

Condition (Stochastic Data Selection)

Data is selected stochastically by drawing from a normalized criterion $\frac{\exp(c(x,\theta_t))}{\int_{x'} \exp(c(x',\theta_t))d\mu(x)}$.

Conditions for Convergence II (ctd.)

Condition (Continuous Selection Criterion)

It holds for the decision criterion $c : \mathcal{X} \times \Theta \to \mathbb{R}$ *in the decision problem* $(\Theta, \mathbb{A}, \ell_{\theta_t})$ *of selecting features to be added to the sample that* $\nabla_x c(x, \theta)$ *and* $\nabla_\theta c(x, \theta)$ *are bounded from above.*

Condition (Linear Selection Criterion)

The decision criterion $c : \mathcal{X} \times \Theta \to \mathbb{R}$ *in* $(\Theta, \mathbb{A}, \ell_{\theta_t})$ *is linear in* x *and Lipschitz-continuous in* θ *with a Lipschitz constant* L_c *that is independent of* x*.*

Conditions for Convergence III (ctd.)

Condition (Soft Labels Prediction)

The prediction function $\hat{y} : \mathcal{X} \times \Theta \longrightarrow \{0,1\}$ on bounded $\mathcal{X}$ gives rise to a non-degenerate distribution of $Y \mid X$ for any θ such that we can consider soft label predictions $p : \mathcal{X} \times \Theta \longrightarrow [0,1]$ with $p(x,\theta) = \sigma(g(X,\theta))$ with $\sigma : \mathbb{R} \longrightarrow [0,1]$ a sigmoid function. Further assume that the loss is jointly smooth in these predictions. That is, $\nabla_p \ell(y, p(x,\theta))$ exists and is Lipschitz-continuous in x and θ.

We can interpret p as $P_\theta(Y \mid X = x)$. In other words, soft labels in the form of probability distributions are available.

References

- Arazo, Eric et al. (2020). "Pseudo-labeling and confirmation bias in deep semi-supervised learning". In: *2020 International Joint Conference on Neural Networks*. IEEE, pp. 1–8.
- Dietrich, Stefan, Julian Rodemann, and Christoph Jansen (2024). "Semi-Supervised Learning guided by the Generalized Bayes Rule under Soft Revision". In: *arXiv preprint arXiv:2405.15294*.
- Fournier, Nicolas and Arnaud Guillin (2015). "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probability theory and related fields* 162.3, pp. 707–738.
- Hastie, Trevor, Robert Tibshirani, and Jerome H Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

References

Hüllermeier, Eyke and Weiwei Cheng (2015). "Superset learning based on generalized loss minimization". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 260–275.

Jansen, C., M. Nalenz, et al. (2023). "Statistical comparisons of classifiers by generalized stochastic dominance". In: *Journal of Machine Learning Research* 24, pp. 1–37.

Jansen, C., G. Schollmeyer, and T. Augustin (2018). "Concepts for decision making under severe uncertainty with partial ordinal and partial cardinal preferences". In: *International Journal of Approximate Reasoning* 98, pp. 112–131.

— (2023). "Multi-target decision making under conditions of severe uncertainty". In: *Modeling Decisions for Artificial Intelligence*. Ed. by V. Torra and Y. Narukawa. Springer, pp. 45–57.

References

Jansen, C., G. Schollmeyer, H. Blocher, et al. (2023). “Robust statistical comparison of random variables with locally varying scale of measurement”. In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*. Ed. by R. Evans and I. Shpitser. Vol. 216. Proceedings of Machine Learning Research. PMLR, pp. 941–952.

Jansen, Christoph et al. (2024). “Statistical Multicriteria Benchmarking via the GSD-Front”. In: *arXiv preprint arXiv:2406.03924*.

Lang, Hunter, Aravindan Vijayaraghavan, and David Sontag (2022). “Training subset selection for weak supervision”. In: *Advances in Neural Information Processing Systems* 35, pp. 16023–16036.

References

- Lee, Dong-Hyun et al. (2013). "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". In: *Workshop on challenges in representation learning, International Conference on Machine Learning*. Vol. 3, p. 896.
- Li, Shuangshuang et al. (2020). "Pseudo-label selection for deep semi-supervised learning". In: *2020 IEEE International Conference on Progress in Informatics and Computing (PIC)*. IEEE, pp. 1–5.
- Liu, Wei et al. (2023). "What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning". In: *arXiv preprint arXiv:2312.15685*.
- Nalenz, Malte, Julian Rodemann, and Thomas Augustin (2024). "Learning de-biased regression trees and forests from complex samples". In: *Machine Learning*, pp. 1–20.

References

Pooladzandi, Omead, David Davini, and Baharan Mirzasoleiman (2022). "Adaptive second order coresets for data-efficient machine learning". In: *International Conference on Machine Learning*. PMLR, pp. 17848–17869.

Rizve, Mamshad Nayeem et al. (2020). "In Defense of Pseudo-Labeling: An Uncertainty-Aware Pseudo-label Selection Framework for Semi-Supervised Learning". In: *International Conference on Learning Representations, 2020*.

Rodemann, Julian (2023). "Pseudo Label Selection is a Decision Problem". In: *Proceedings of the 46th German Conference on Artificial Intelligence*. Springer.

— (2024). "Bayesian Data Selection". In: *arXiv preprint arXiv:2406.12560*. 5th Workshop on Data-Centric Machine Learning Research (DMLR) at ICML 2024.

References

- Rodemann, Julian and Augustin (2024). "Imprecise Bayesian optimization". In: *Knowledge-Based Systems* 300, p. 112186. ISSN: 0950-7051. DOI: `https://doi.org/10.1016/j.knosys.2024.112186`.
- Rodemann, Julian and Thomas Augustin (2021). "Accounting for imprecision of model specification in Bayesian optimization". In: *Poster presented at International Symposium on Imprecise Probabilities (ISIPTA)*.
- — (2022). "Accounting for Gaussian process imprecision in Bayesian optimization". In: *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM)*. Springer, pp. 92–104.
- Rodemann, Julian and Hannah Blocher (2024). "Partial Rankings of Optimizers". In: *International Conference on Learning Representations (ICLR), Tiny Papers Track*.

References

- Rodemann, Julian, Federico Croppi, et al. (2024). "Explaining Bayesian Optimization by Shapley Values Facilitates Human-AI Collaboration". In: *arXiv preprint arXiv:2403.04629*.
- Rodemann, Julian, Sebastian Fischer, et al. (2022). "Not All Data Are Created Equal: Lessons From Sampling Theory For Adaptive Machine Learning". In: *International Conference on Statistics and Data Science (ICSDS) by the Institute of Mathematical Statistics (IMS)*.
- Rodemann, Julian, Jann Goschenhofer, et al. (2023). "Approximately Bayes-optimal pseudo-label selection". In: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*. Vol. 216. Proceedings of Machine Learning Research. PMLR, pp. 1762–1773.

References

- Rodemann, Julian, Christoph Jansen, et al. (2023). “In all likelihoods: Robust selection of pseudo-labeled data”. In: *International Symposium on Imprecise Probability: Theories and Applications*. PMLR, pp. 412–425.
- Rodemann, Julian, Dominik Kreiss, et al. (2022). “Levelwise Data Disambiguation by Cautious Superset Classification”. In: *International Conference on Scalable Uncertainty Management*. Springer, pp. 263–276.
- Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.
- Stolz, Bernadette J (2023). “Outlier-robust subsampling techniques for persistent homology”. In: *Journal of Machine Learning Research* 24.90, pp. 1–35.

References

Yin, Zimo et al. (2024). "Embrace sustainable AI: Dynamic data subset selection for image classification". In: *Pattern Recognition*, p. 110392.