Recent Results on Algorithmic Fairness and Meta-Learning

Massimiliano Pontil

Computational Statistics and Machine Learning
Istituto Italiano di Tecnologia
and
Department of Computer Science
University College London

4th Annual Machine Learning in the Real World workshop (MLiTRW)
Criteo AI Lab, Paris, October 2, 2019

Plan

- Fair empirical risk minimization
- Using labeled and unlabeled data
- Multi-task approach
- Learning fair representations
- Online meta-learning

Algorithmic fairness

- **Aim:** ensure that learning algorithms do not treat subgroups in the population "unfairly"
- **How:** impose "fairness" constraints (different notions)
- **Difficulty:** study computationally efficient algorithms with statistical guarantees w.r.t. both the risk and the fairness measure

Binary classification setting: let μ be a prob. distribution on $\mathcal{X} \times \mathcal{S} \times \{-1, +1\}$, where $\mathcal{S} = \{a, b\}$ is the sensitive variable set. We wish to find a solution f^* of

$$\min_{f \in \mathcal{F}} \left\{ \mathbb{P}\big(f(X, S) \neq Y\big) \; s.t. \; \text{"}f \text{ is fair"} \right\}$$

Fairness constraints

(see e.g. [Hardt et al., 2016, Zafar et al., 2017])

- Equal opportunity (EO): $\mathbb{P}\big(f(X,S){>}0|Y{=}1, S{=}a\big) = \mathbb{P}\big(f(X,S){>}0|Y{=}1, S{=}b\big)$

- Equalized odds (EOd): $f(X,S)$ and S are conditionally independent given Y, i.e.

$$\mathbb{P}\big(f(X,S){>}0|Y{=}y, S{=}a\big) = \mathbb{P}\big(f(X,S){>}0|Y{=}y, S{=}b\big), \quad y \in \{-1,1\}$$

- Demographic parity (DP): $\mathbb{P}\big(f(X,S){>}0|S{=}a\big) = \mathbb{P}\big(f(X,S){>}0|S{=}b\big)$

- We may also loose each constraint by requiring the l.h.s. to be close to the r.h.s.

Statistical learning setting

- Let $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ be a loss function and let L be the associated risk:

$$L(f) = \mathbb{E}[\ell(f(X), Y)], \quad \text{for } f : \mathcal{X} \to \mathcal{Y}$$

- Conditional risk of f for the positive class in group s:

$$L^{+,s}(f) = \mathbb{E}[\ell(f(X), Y) | Y = 1, S = s]$$

- We relax the fairness constraint by using a loss function in place of the 01-loss and introduce a parameter $\epsilon \in [0, 1]$. For EO, we obtain

$$\min_{f \in \mathcal{F}} \left\{ L(f) \;\; s.t. \;\; \left| L^{+,a}(f) - L^{+,b}(f) \right| \leq \epsilon \right\} \tag{1}$$

Fair empirical risk minimization (FERM)

[Donini et al. NeurIPS 2018]

- Distribution μ is unknown and we only have a data sequence $(x_i, s_i, y_i)_{i=1}^n$ sampled independently from μ. We then consider the empirical problem

$$\min_{f \in \mathcal{F}} \left\{ \hat{L}(f) \;\; s.t. \;\; \left| \hat{L}^{+,a}(f) - \hat{L}^{+,b}(f) \right| \leq \hat{\epsilon} \right\} \tag{2}$$

  where $\hat{\epsilon}$ is a parameter linked to ϵ

- Empirical risk $\hat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$
- Empirical risk for the positive samples in group g: $\hat{L}^{+,g}(f) = \frac{1}{n^{+,g}} \sum_{i \in \mathcal{I}^{+,g}} \ell(y_i, f(x_i))$ with $\mathcal{I}^{+,g} = \{i : y_i{=}1, s_i{=}g\}$ and $n^{+,g} = |\mathcal{I}^{+,g}|$, $g \in \{a, b\}$

Statistical analysis of FERM

We say a class of functions $\mathcal{F}$ is learnable (wrt. loss ℓ) if:

$$\sup_{f \in \mathcal{F}} \left| L(f) - \hat{L}(f) \right| \leq B(\delta, n, \mathcal{F}), \quad \text{with} \lim_{n \to \infty} B(\delta, n, \mathcal{F}) = 0$$

Proposition 1. Let $\delta \in (0, 1)$. If $\mathcal{F}$ is learnable f^* solves (1) and $\hat{f}$ solves (2) with $\hat{\epsilon} = \epsilon + \sum_{g \in \{a,b\}} B(\delta, n^{+,g}, \mathcal{F})$ then with prob. $\geq 1 - 6\delta$ it holds simultaneously:

$$L(\hat{f}) - L(f^*) \leq 2B(\delta, n, \mathcal{F})$$

$$\left| L^{+,a}(\hat{f}) - L^{+,b}(\hat{f}) \right| \leq \epsilon + 2 \sum_{g \in \{a,b\}} B(\delta, n^{+,g}, \mathcal{F})$$

Implications of the bound

- Bound implies that a solution $\hat{f}$ of (2) is close to a solution of f^* of (1) both in terms of the risk and fairness constraint
- But how do we find $\hat{f}$? We would like to solve problem (2) for the hard (misclassification) loss:

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} 1\{f(x_i) \neq y_i\} \tag{3}$$
$$\left| \hat{\mathbb{P}}\{f(x) > 0 | y = 1, s = a\} - \hat{\mathbb{P}}\{f(x) > 0 | y = 1, s = b\} \right| \leq \epsilon$$

- We propose to replace the hard loss in the risk with the (larger) hinge loss, and the hard loss in the fairness constraint with a linear loss

Fair learning with kernels

- Linear model $f(\cdot) = \langle w, \phi(\cdot) \rangle$, with $\phi : \mathcal{X} \to \mathbb{H}$ a kernel-induced feature map
- For the linear loss, the fairness constraint takes the form $\left| \langle w, u_a - u_b \rangle \right| \leq \hat{\epsilon}$, where u_g is the barycenter of the positive points in group g:

$$u_g = \frac{1}{n^{+,g}} \sum_{i:\in \mathcal{I}^{+,g}} \phi(x_i), \quad g \in \{a, b\}$$

- We consider the regularized empirical risk minimization problem

$$\min_{w \in \mathbb{H}} \sum_{i=1}^{n} \ell(y_i \langle w, \phi(x_i) \rangle) + \lambda \|w\|^2 \quad \text{s.t.} \quad \left| \langle w, u_a - u_b \rangle \right| \leq \hat{\epsilon} \qquad \lambda > 0$$

Form of the optimal classifier

[Chzhen et al. NeurIPS 2019]

Proposition. Let $\eta(x,s) = \mathbb{E}[Y \mid X = x, S = s]$ be the regression function. If for each $s \in \{0,1\}$ the mapping $t \mapsto \mathbb{P}(\eta(X,S) \leq t \mid S = s)$ is continuous on $(0,1)$, then an optimal classifier f^* can be obtained for all $(x,s) \in \mathbb{R}^d \times \{a,b\}$ as

$$f_\theta(x,a) = \mathbf{1}_{\{1 \leq \eta(x,a)(2 - \frac{\theta}{\mathbb{P}(Y=1,S=a)})\}}, \quad f_\theta(x,b) = \mathbf{1}_{\{1 \leq \eta(x,b)(2 + \frac{\theta}{\mathbb{P}(Y=1,S=b)})\}}$$

where $\theta \in [0,2]$ solves the equation

$$\frac{\mathbb{E}_{X|S=a}[\eta(X,a) f_\theta(X,a)]}{\mathbb{P}(Y=1 \mid S=a)} = \frac{\mathbb{E}_{X|S=b}[\eta(X,b) f_\theta(X,b)]}{\mathbb{P}(Y=1 \mid S=b)} \,.$$

- Similar result when S is not included as a predictor

Leveraging labeled and unlabeled

- FERM leaves open the question of designing a computationally efficient and statistically consistent estimator for problem (*)
- Alternative method: estimate η from a labeled sample and θ from an independent unlabeled sample by minimizing the empirical difference of equal opportunity (DEO)

$$\hat{\Delta}(f,\mu) = \left| \frac{\hat{\mathbb{E}}_{X|S=a}\hat{\eta}(X,a)f_\theta(X,a)}{\hat{\mathbb{E}}_{X|S=a}\hat{\eta}(X,a)} - \frac{\hat{\mathbb{E}}_{X|S=b}\hat{\eta}(X,b)f_\theta(X,b)}{\hat{\mathbb{E}}_{X|S=b}\hat{\eta}(X,b)} \right|$$

Theorem (informal). If $\hat{\eta} \to \eta$ as $n \to \infty$, under mild additional assumptions the proposed estimator is consistent w.r.t. both accuracy and fairness:

$$\lim_{n,N\to\infty} \mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}[\Delta(\hat{f},\mu)] = 0 \quad \text{and} \quad \lim_{n,N\to\infty} \mathbb{E}_{(\mathcal{D}_n,\mathcal{D}_N)}[\mathcal{R}(\hat{f})] \leq \mathcal{R}(f^*)$$

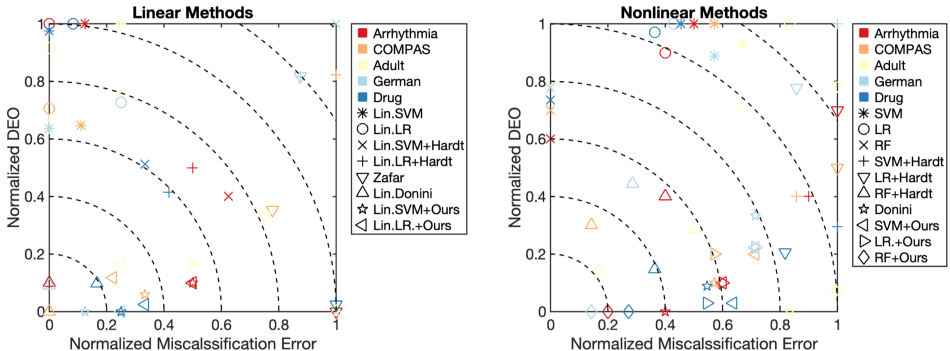
Modified validation procedure

- In experiments, we employ a two steps 10-fold CV procedure:
  - Step 1: shortlist all hyperparameters with accuracy above a certain percentage (we choose 90%) of the best accuracy
  - Step 2, from the list, select the hyperparameter with highest fairness (i.e. lowest DEO)
- We compare:
  - Naïve SVM, validated with a standard nested 10-fold cross validation
  - SVM with the novel validation procedure
  - The method by [Hardt et al., 2016] applied to the best SVM
  - The method [Zafar et al., 2017] (code provided by the authors for the linear case*)

*Python code: `https://github.com/mbilalzafar/fair-classification`

Experiments

Comparison between different methods. DEO is normalized in $[0, 1]$ column-wise. The closer a point is to the origin, the better the result



The proposed methods slightly decrease accuracy while greatly improving in the fairness measure

Code: `https://github.com/jmikko/fair_ERM`

Taking advantage of multitask learning

[Oneto et al. AIES 2019]

- We consider group specific models: $f(x, s) = \langle w_s, x \rangle$ and a multitask learning (MTL) formulation

$$\min_{w_1,\dots w_k \in \mathbb{H}} \sum_{s=1}^{k} \hat{L}_s(w_s) + \frac{\lambda}{k} \sum_{s=1}^{k} \|w_s - w_0\|^2 + (1-\lambda)\|w_0\|^2$$

- Regularization around a common mean encourages task similarities
- We impose additional (linearized) fairness constraints on f and the common mean

Left: Shared model trained with MTL, with fairness constraint, and no sensitive feature in the predictors vs. the group specific models trained with MTL, with fairness constraint

Right: The latter models vs. the same models when the sensitive feature is predicted via random forest

P - Pred. or not sens. feat.

D - Group specific models or not

F - Fair. const. is active or not

S - Using or not pred. sens. feat.

	P	D	F	S	MTL ACC	MTL DEOd
Adult Dataset						
G	0	0	1	0	82.0	0.06
	0	1	1	1	88.3	0.03
R	0	0	1	0	82.8	0.01
	0	1	1	1	90.6	0.03
G+R	0	0	1	0	83.5	0.04
	0	1	1	1	90.3	0.05
COMPAS Dataset						
G	0	0	1	0	75.7	0.03
	0	1	1	1	82.1	0.06
R	0	0	1	0	82.6	0.03
	0	1	1	1	90.2	0.03
G+R	0	0	1	0	83.4	0.05
	0	1	1	1	90.3	0.05

	P	D	F	S	MTL ACC	MTL DEOd
Adult Dataset						
G	0	1	1	1	88.3	0.03
	1	1	1	1	87.6	0.01
R	0	1	1	1	90.6	0.03
	1	1	1	1	89.4	0.01
G+R	0	1	1	1	90.3	0.05
	1	1	1	1	89.3	0.01
COMPAS Dataset						
G	0	1	1	1	82.1	0.06
	1	1	1	1	81.3	0.01
R	0	1	1	1	90.2	0.03
	1	1	1	1	89.2	0.01
G+R	0	1	1	1	90.3	0.05
	1	1	1	1	89.3	0.01

Learning fair representations

[Oneto et al. Arxiv 2019]

- Now consider demographic parity: $\mathbb{P}(f(x) = 1|S = 0) = \mathbb{P}(f(x) = 1|S = 1)$
- Suppose $f(x) = g(h(x))$. If representation $h : \mathcal{X} \to \mathbb{R}^r$ is fair in the following sense

$$\mathbb{P}(h(x) \in C|S = a) = \mathbb{P}(h(x) \in C|S = b), \quad \forall C \in \mathbb{R}^r$$

  then f is fair as well

- We relax this by requiring that both distributions have the same means.
  We let $c(\mathbf{z})$ the difference of the empirical means from a dataset $\mathbf{z}$
- We use multiple tasks to learn h. We illustrate the approach in the linear case, $h(x) = A^\top x$, and $f(x) = b^\top h(x)$:

$$\min_{A,B} \left\{ \frac{1}{Tm} \sum_{t=1}^{T} \sum_{i=1}^{n} \left(y_{t,i} - \langle b_t, A^\top x_{t,i} \rangle\right)^2 + \frac{\lambda}{2} \|A\|_F \|B\|_F \;\middle|\; A^\top c(\mathbf{z}_t) = 0,\ 1 \le t \le T \right\}$$

Learning fair representations (cont.)

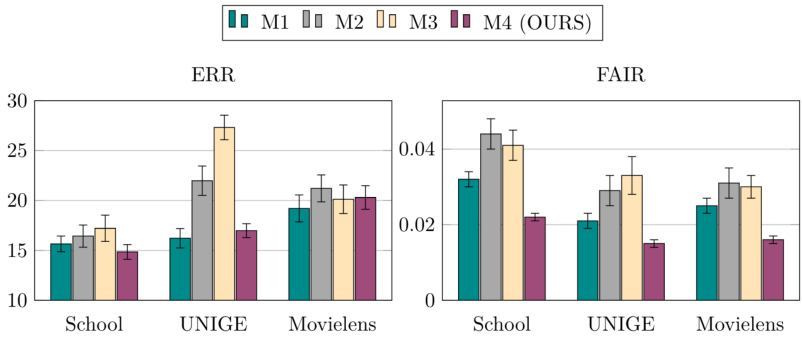
Theorem. Let A solve the above problem and $\|A\|_F = 1$. Let tasks $\mu_1, \ldots, \mu_T$ be i.i.d. from a meta-distribution ρ. Then, with probability at least $1 - \delta$, the average risk of the algorithm with representation A run on a random task is upper bounded

$$\frac{1}{Tn} \sum_{t=1}^{T} \sum_{i=1}^{n} \left(y_{t,i} - \langle b_t, A^\top x_{t,i} \rangle \right)^2 + O\left(\frac{1}{\lambda} \sqrt{\frac{\|\hat{C}\|_\infty}{n}} \right) + O\left(\sqrt{\frac{\ln \frac{1}{\delta}}{T}} \right)$$

and the linearized fairness constraint is bounded as

$$\mathbb{E}_{\mu \sim \rho} \mathbb{E}_{\mathbf{z} \sim \mu^n} \|Ac(\mathbf{z})\|^2 \leq \frac{1}{T} \sum_{t=1}^{T} \|Ac(\mathbf{z}_t)\|^2 + 96 \frac{\ln \frac{8T^2}{\delta}}{T} + 6 \sqrt{\frac{\|\hat{\Sigma}\|_\infty \ln \frac{8T^2}{\delta}}{T}}$$

Experiments



M1: Standard MTL with the fairness constraints on the outputs
M2: Feed-forward neural network (FFNN) with adversarially generated representation *[Madras et al. ICML 2018]*
M3: Similar to M2 but with different loss function *[Edwards &Storkey, ICLR 2016]*

From MTL to meta-learning[†]

From a sequence of tasks find an algorithm which works well on unseen similar tasks

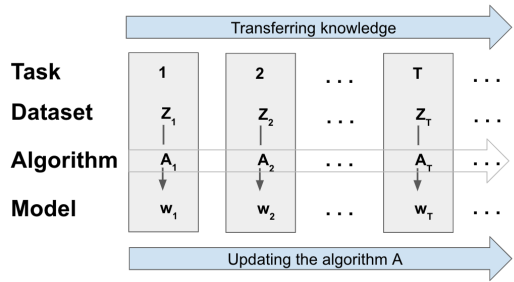
task	11111	2222	333333	···
data	12345	1234	123456	···

- Previous work mainly focused on the batch statistical setting
  [Baxter, 2000, Maurer, 2009, Pentina and Lampert, 2014, Maurer et al., 2016]
- Recent interest on online meta-learning:
  - **Online-within-online:** both tasks and within-task data arrive online
    [Alquier et al., 2017, Denevi et al., 2019, Khodak et al., 2019]
  - **Online-within-batch:** tasks arrive online, their datasets in one batch
    [Denevi et al., 2018a, Denevi et al., 2018b, Finn et al., 2019, Bullins et al., 2019]
- Also recent interest on meta-learning with deep neural networks, e.g.
  [Ravi and Larochelle, 2017, Finn et al., 2017, Franceschi et al., 2018]

[†] Equivalent terminology: learning-to-learn or lifelong learning

Meta-algorithm

A model for each task is learned by an inner algorithm, which is updated by a meta-algorithm as the tasks are sequentially observed



- Desiderata: memory and time efficient, and supported by learning guarantees
- Difficulty: lack of a convex meta-objective

Statistical and non-statistical settings

Let $Z_t = (x_{t,i}, y_{t,i})_{i=1}^n$ be the training sequence for the t-th task and let $\mathbf{Z} = (Z_t)_{t=1}^T$ be the meta-sequence. We consider two settings[‡]

- Statistical setting [Baxter, 2000, Maurer, 2009]: the tasks are sampled from a meta-distribution ρ and we wish to bound the average excess risk

$$\mathbb{E}_{\mathbf{Z}} \mathbb{E}_{\mu\sim\rho} \mathcal{E}_\mu(A(Z)) = \mathbb{E}_{\mathbf{Z}} \left[\mathbb{E}_{\mu\sim\rho} \left[\mathbb{E}_{Z\sim\mu^n} \mathcal{R}_\mu\big(A(Z)\big) - \min_{w\in\mathbb{R}^d} \mathcal{R}_\mu(w) \right] \right]$$

- Non-statistical setting: we wish to bound the normalized regret across the tasks

$$\mathit{regret}(A_1, ..., A_T) = \frac{1}{T} \sum_{t=1}^T \left\{ \frac{1}{n} \sum_{i=1}^n \ell\big(\langle x_{t,i}, w_{t,i}\rangle, y_{t,i}\big) - \min_{w\in\mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \ell\big(\langle x_{t,i}, w\rangle, y_{t,i}\big) \right\}$$

[‡] See [Alquier et al., 2017] for a discussion

Regularizaton around a common mean – learning guarantees

[Denevi et al. ICML 2019; Denevi et al. NeurIPS 2019]

We assume $\ell(\cdot, y)$ L-Lipschitz for any $y \in \mathcal{Y}$ and the inputs are bounded. Let w_μ be the minimizer of the true risk for task μ

$$V_\rho(\theta) = \frac{1}{2}\mathbb{E}_{\mu\sim\rho}\|w_\mu - \theta\|_2^2 \qquad \theta_\rho = \operatorname*{argmin}_{\theta\in\Theta} V_\rho(\theta) = \mathbb{E}_{\mu\sim\rho} w_\mu$$

- **Our method** (from Thm. 2, tuning of λ and η)

$$\mathbb{E}_{\mathbf{Z}}\ \mathbb{E}_{\mu\sim\rho}\ \mathcal{E}_\mu(A_{\bar\theta}) \leq\ \mathcal{O}\left(\sqrt{\frac{V_\rho(\theta_\rho)}{n}} + \sqrt{\frac{1}{T}}\right)$$

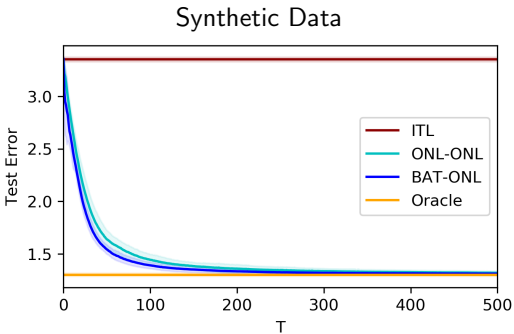
- **Best algorithm** $\theta = \theta_\rho$

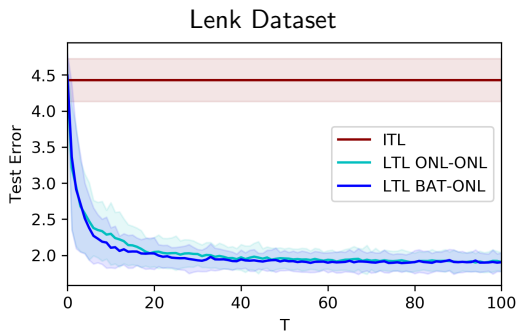
$$\mathbb{E}_{\mu\sim\rho}\ \mathcal{E}_\mu(A_\theta) \leq\ \mathcal{O}\left(\sqrt{\frac{V_\rho(\theta_\rho)}{n}}\right)$$

- **Indep. task learning (ITL)** $\theta = 0$

$$\mathbb{E}_{\mu\sim\rho}\ \mathcal{E}_\mu(A_\theta) \leq\ \mathcal{O}\left(\sqrt{\frac{V_\rho(0)}{n}}\right)$$

Experiment





Averaged test performance of different methods on synthetic (**Left**) and the Lenk dataset (**Right**) as the number of training tasks incrementally increases.

We are hiring!

Postdoc/Researcher positions at Istituto Italiano diTecnologia in Genoa to work with me

Send me an email if interested: massimiliano.pontil@iit.it
More info: `http://tinyurl.com/MLPostDocIIT2019`

References I

Alquier, P., Mai, T. T., and Pontil, M. (2017).
Regret bounds for lifelong learning.
In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 261–269.

Baxter, J. (2000).
A model of inductive bias learning.
J. Artif. Intell. Res., 12(149–198):3.

Bullins, B., Hazan, E., Kalai, A., and Livni, R. (2019).
Generalize across tasks: Efficient algorithms for linear representation learning.
In *Algorithmic Learning Theory*, pages 235–246.

Denevi, G., Ciliberto, C., Grazzi, R., and Pontil, M. (2019).
Learning-to-learn stochastic gradient descent with biased regularization.
arXiv preprint arXiv:1903.10399.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. (2018a).
Incremental learning-to-learn with statistical guarantees.
In *Proc. 34th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Denevi, G., Ciliberto, C., Stamos, D., and Pontil, M. (2018b).
Learning to learn around a common mean.
In *Advances in Neural Information Processing Systems*, pages 10190–10200.

References II

Finn, C., Abbeel, P., and Levine, S. (2017).
Model-agnostic meta-learning for fast adaptation of deep networks.
In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. (2019).
Online meta-learning.
arXiv preprint arXiv:1902.08438.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. (2018).
Bilevel programming for hyperparameter optimization and meta-learning.
In *International Conference on Machine Learning*, PMLR 80, pages 568–1577.

Hardt, M., Price, E., and Srebro, N. (2016).
Equality of opportunity in supervised learning.
In *Advances in Neural Information Processing Systems*.

Khodak, M., Balcan, M.-F., and Talwalkar, A. (2019).
Provable guarantees for gradient-based meta-learning.
arXiv preprint arXiv:1902.10644.

Maurer, A. (2009).
Transfer bounds for linear feature learning.
Machine Learning, 75(3):327–350.

References III


Maurer, A., Pontil, M., and Romera-Paredes, B. (2016).
The benefit of multitask representation learning.
The Journal of Machine Learning Research, 17(1):2853–2884.


Pentina, A. and Lampert, C. (2014).
A PAC-Bayesian bound for lifelong learning.
In *International Conference on Machine Learning*, pages 991–999.


Ravi, S. and Larochelle, H. (2017).
Optimization as a model for few-shot learning.
In *I5th International Conference on Learning Representations*.


Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017).
Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment.
In *International Conference on World Wide Web*.