

COSC 594/650

Constrained Empirical Risk Minimization

Dongsheng Ding

dongshed@utk.edu

Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville

Recap: empirical risk minimization

Data: $(x, y) \sim \mathbb{P}^*$, unknown $\mathbb{P}^*$

Training data: $(x_1, y_1), \dots, (x_m, y_m)$, drawn from $\mathbb{P}^*$

■ THE ERROR WE CARE ABOUT, AND THE ERROR WE CAN MEASURE

$$L(h) = \mathbb{E}_{\mathbb{P}^*} [\ell((x, y), h)], \quad \hat{L}(h) = \frac{1}{m} \sum_{i=1}^m \ell((x_i, y_i), h)$$

■ EMPIRICAL RISK MINIMIZATION

$$\hat{h}^* \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{L}(h)$$

★ $L(\hat{h}^*)$ is close to $L(h^*)$ with high probability – Lecture 17

One loss: accuracy is the only goal

QUESTION: Can we learn under requirements?

Requirements versus goals

Goal: the model *should* be accurate – an objective

Requirement: the model *shall* be fair, robust, ... – a constraint

Requirement	Written as a bound on a risk
Fairness	$\mathbb{E} [h(x) \mid a = 0] - \mathbb{E} [h(x) \mid a = 1] \leq c$
Robustness	$\mathbb{E} [\max_{\ v\ \leq r} \ell((x + v, y), h)] \leq c$
Every client j	$\mathbb{E}_{\mathbb{P}_j} [\ell((x, y), h)] \leq c_j$

Requirements: constraints on risks, not terms in the loss

Constrained learning

Objective: $L_0(h) = \mathbb{E}_{\mathbb{P}^*} [\ell_0((x, y), h)]$

Requirements: $L_j(h) = \mathbb{E}_{\mathbb{P}_j} [\ell_j((x, y), h)], \quad j = 1, \dots, q$

■ THE PROBLEM, AND ITS SOLUTION h^*

$$\begin{aligned} p^* &= \underset{h \in \mathcal{H}}{\text{minimize}} && L_0(h) \\ &\text{subject to} && L_j(h) \leq c_j, \quad j = 1, \dots, q \end{aligned}$$

- ★ $\mathbb{P}_j$ can differ from $\mathbb{P}^*$ – e.g., the data of one group
- ★ $q = 0$ – the expected risk minimization of Lecture 16

Constrained learning: minimize one risk, subject to bounds on others

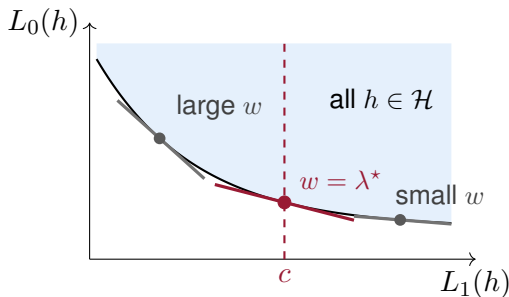
Penalties versus requirements

Penalty: $h_w \in \operatorname{argmin}_{h \in \mathcal{H}} L_0(h) + w L_1(h)$, fixed weight $w \geq 0$

■ WHICH WEIGHT?

- ★ small w – $L_1(h_w) > c$, violates
- ★ large w – $L_1(h_w) < c$, wasteful
- ★ $w = \lambda^*$ – exactly $L_1(h_w) = c$

■ BUT λ^* DEPENDS ON $\mathbb{P}^*$ AND c



Penalty: the right weight is λ^* , unknown in advance

Constrained empirical risk minimization

Training data: m samples from $\mathbb{P}^*$, and m samples from each $\mathbb{P}_j$

$$\begin{aligned} \hat{p}^* = \underset{h \in \mathcal{H}}{\text{minimize}} \quad & \hat{L}_0(h) \\ \text{subject to} \quad & \hat{L}_j(h) \leq c_j, \quad j = 1, \dots, q \end{aligned}$$

■ TWO QUESTIONS ABOUT ITS SOLUTION $\hat{h}^*$

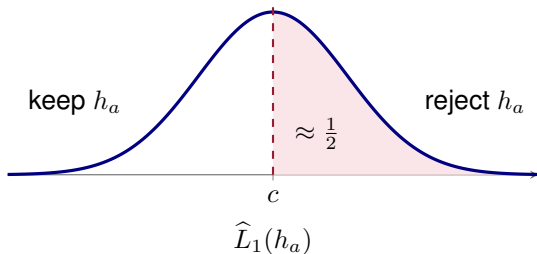
- ★ statistical – on new data, $L_0(\hat{h}^*) \approx p^*$ and $L_j(\hat{h}^*) \leq c_j$?
- ★ computational – a nonconvex feasible set, with no projection

Constrained ERM: meet $\hat{L}_j(h) \leq c_j$, and hope $\hat{h}^*$ meets $L_j(h) \leq c_j$

Constrained ERM can fail

Two hypotheses: $\mathcal{H} = \{h_a, h_b\}$, h_a is more accurate: $L_0(h_a) < L_0(h_b)$

Requirement: $L_1(h_a) = c$ (active), $L_1(h_b) < c \implies h^* = h_a$



★ rejected with probability $\approx 1/2$, for every m – then $\hat{h}^* = h_b$

Active requirement: the sample check is a coin flip

The empirical dual problem

Lagrangian: $\hat{\mathcal{L}}(h, \lambda) = \hat{L}_0(h) + \sum_{j=1}^q \lambda_j \left(\hat{L}_j(h) - c_j \right), \quad \lambda \geq 0$

■ DUAL FUNCTION

$$\hat{g}(\lambda) = \min_{h \in \mathcal{H}} \hat{\mathcal{L}}(h, \lambda)$$

- ★ concave in λ – for any $\mathcal{H}$ and any losses
- ★ fixed λ – an unconstrained ERM, with weights λ_j on the requirements

■ DUAL PROBLEM

$$\hat{d}^{\star} = \max_{\lambda \geq 0} \hat{g}(\lambda) = \hat{g}(\hat{\lambda}^{\star}) \leq \hat{p}^{\star}$$

The dual: one ERM for each λ , and a concave problem in λ

Dual ascent for constrained learning

Parameters: dual step size α , an ERM solver such as SGD

$\lambda_0 = 0$; **for** $k = 0, 1, 2, \dots$

$h_k \approx \operatorname{argmin}_{h \in \mathcal{H}} \hat{\mathcal{L}}(h, \lambda_k)$ (a few epochs of SGD, started at h_{k-1})

$\lambda_{k+1} = \left[\lambda_k + \alpha \left(\hat{L}(h_k) - c \right) \right]_+, \quad \hat{L} = (\hat{L}_1, \dots, \hat{L}_q)$

■ DUAL UPDATE (LECTURE 14)

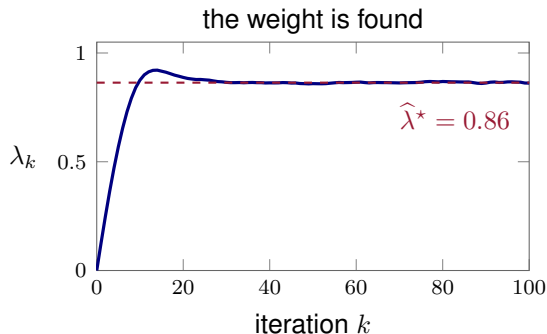
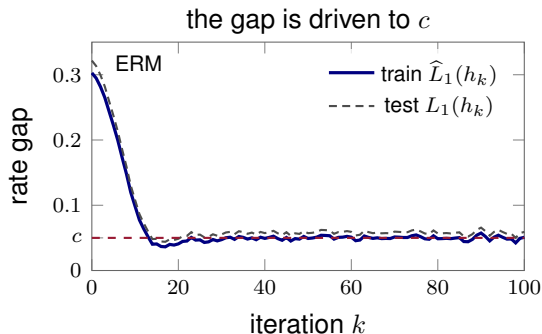
- ★ violation raises λ_j , slack lowers it – floor at zero
- ★ $\lambda_k \rightarrow \hat{\lambda}^*$ – slide 5's weight λ^* , estimated from data

Dual ascent: the weights tune themselves to the requirements

Experiment: a fairness requirement

Data: $m = 2000$, two groups $a \in \{0, 1\}$, logistic regression (Lecture 16)

Requirement: $L_1(h) = \mathbb{E}[h(x) | a = 0] - \mathbb{E}[h(x) | a = 1] \leq c = 0.05$



No tuning: dual ascent finds the weight and meets the requirement

How good is the empirical dual?

True dual: $\mathcal{L}, g, d^*, \lambda^*$ – as on slide 8, with the true risks L_j in place of $\widehat{L}_j$

Enough data: $|\widehat{L}_j(h) - L_j(h)| \leq \zeta \sim 1/\sqrt{m}$ for all $h \in \mathcal{H}$ and all j , w. p. $1 - \delta$

■ DUAL GAP

$$|d^* - \widehat{d}^*| \leq (1 + \Lambda) \zeta, \quad \Lambda = \max \left\{ \|\lambda^*\|_1, \|\widehat{\lambda}^*\|_1 \right\}$$

★ strong duality $p^* = d^*$ (Lecture 13) – the same bound for $|p^* - \widehat{d}^*|$

Prices, not hard checks: small errors in the risks, small error in $\widehat{d}^*$

Proof: one λ at a time

Same weights: $\widehat{\mathcal{L}}$ and $\mathcal{L}$ weight the risks by $1, \lambda_1, \dots, \lambda_q$; the c_j cancel

■ FIX λ AND h

$$|\widehat{\mathcal{L}}(h, \lambda) - \mathcal{L}(h, \lambda)| \leq |\widehat{L}_0(h) - L_0(h)| + \sum_{j=1}^q \lambda_j |\widehat{L}_j(h) - L_j(h)| \leq (1 + \|\lambda\|_1) \zeta$$

■ MINIMIZE OVER h ON BOTH SIDES

$$|\widehat{g}(\lambda) - g(\lambda)| \leq (1 + \|\lambda\|_1) \zeta$$

■ EVALUATE AT THE MAXIMIZERS

$$d^* = g(\lambda^*) \leq \widehat{g}(\lambda^*) + (1 + \|\lambda^*\|_1) \zeta \leq \widehat{d}^* + (1 + \Lambda) \zeta$$

★ the same three steps at $\widehat{\lambda}^* - \widehat{d}^* \leq d^* + (1 + \Lambda) \zeta$

Weighted risks: the error grows by $1 + \|\lambda\|_1$, and no more

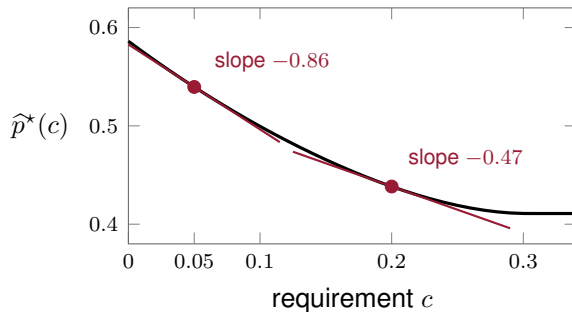
The price of a requirement

Sensitivity (Lecture 13): $\hat{p}^*(c+u) \geq \hat{p}^*(c) - \hat{\lambda}^*u$, $\hat{\lambda}^* = -\frac{d\hat{p}^*}{dc}$

■ FAIRNESS EXPERIMENT

- ★ $c = 0.05 - \hat{\lambda}^* = 0.86$
- ★ $c = 0.20 - \hat{\lambda}^* = 0.47$
- ★ $c \geq 0.31 - \hat{\lambda}^* = 0$, free

■ TIGHTER c , HIGHER PRICE



The multiplier: what a requirement costs, read off the dual

Constrained empirical risk minimization

$$\underset{h \in \mathcal{H}}{\text{minimize}} \quad \widehat{L}_0(h) \quad \text{subject to} \quad \widehat{L}_j(h) \leq c_j, \quad j = 1, \dots, q$$

- REQUIREMENTS ARE CONSTRAINTS ON RISKS
- THE DUAL TURNS THEM INTO WEIGHTED ERM
- THE MULTIPLIERS ARE THE PRICES OF THE REQUIREMENTS

Quantity	Meaning	Available to us
$L_0(h), L_j(h)$	objective, requirements	no, $\mathbb{P}^*$ and $\mathbb{P}_j$ are unknown
$\widehat{\lambda}^*$	price of each requirement	yes, from dual ascent
$ d^* - \widehat{d}^* $	price of the proxy	at most $(1 + \Lambda)\zeta$

Questions?

(**Email: dongshed@utk.edu**)