

COSC 594/650

Stochastic optimization

Dongsheng Ding

dongshed@utk.edu

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

From hypotheses to parameters

Expected risk minimization: $h^* \in \operatorname{argmin}_{h \in \mathcal{H}} L(h)$

Empirical risk minimization: $\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \hat{L}(h)$

■ TWO CHANGES OF NOTATION

- ★ a hypothesis $h \in \mathcal{H}$ is a parameter $\theta \in \mathbb{R}^d$
- ★ an example (x, y) is a sample ξ , drawn from $\mathbb{P}^*$

■ THE SAME TWO PROBLEMS

$$\theta^* \in \operatorname{argmin}_{\theta} \mathbb{E}_{\xi} [\ell(\xi, \theta)] \quad \text{and} \quad \hat{\theta}_m^* \in \operatorname{argmin}_{\theta} \frac{1}{m} \sum_{i=1}^m \ell(\xi_i, \theta)$$

Empirical risk minimization: sample average approximation

The stochastic optimization problem

Population problem: $\underset{\theta}{\text{minimize}} \quad L(\theta) = \mathbb{E}_{\xi} [\ell(\xi, \theta)]$

Optimal value: $L^* = \underset{\theta}{\text{minimize}} L(\theta), \quad \text{attained at } \theta^*$

■ NOT AN ORDINARY OPTIMIZATION PROBLEM

- ★ $\mathbb{P}^*$ is unknown – neither $L(\theta)$ nor $\nabla L(\theta)$ can be evaluated
- ★ even for known $\mathbb{P}^*$, $L(\theta)$ is an integral in high dimension

■ STANDING ASSUMPTIONS

- ★ $\ell(\xi, \cdot)$ convex and differentiable, and θ^* unique

The objective is an expectation, and an algorithm sees it only through **samples**

Sample average approximation

Samples: $\xi_1, \dots, \xi_m$ drawn i.i.d. from $\mathbb{P}^*$

Empirical risk:
$$\hat{L}_m(\theta) = \frac{1}{m} \sum_{i=1}^m \ell(\xi_i, \theta)$$

■ REPLACE THE DISTRIBUTION, KEEP THE PROBLEM

$$\hat{L}_m^* = \underset{\theta}{\text{minimize}} \hat{L}_m(\theta) \quad \text{and} \quad \hat{\theta}_m^* \in \underset{\theta}{\text{argmin}} \hat{L}_m(\theta)$$

■ WHAT IS RANDOM

★ $\hat{L}_m$, $\hat{L}_m^*$, and $\hat{\theta}_m^*$ are random – L , L^* and θ^* are not

Sample average approximation is a problem, and not an algorithm

QUESTION: How close is $\hat{\theta}_m^*$ to θ^* , and how fast?

One parameter at a time

Fix θ : chosen before the samples are drawn

■ LAW OF LARGE NUMBERS

$$\widehat{L}_m(\theta) = \frac{1}{m} \sum_{i=1}^m \ell(\xi_i, \theta) \longrightarrow L(\theta) \quad \text{almost surely}$$

★ $\ell(\xi_1, \theta), \dots, \ell(\xi_m, \theta)$ are i.i.d. with mean $L(\theta)$

■ WHY THIS IS NOT ENOUGH

- ★ $\widehat{\theta}_m^*$ is chosen after the samples – it is not a fixed θ
- ★ the minimizer of a limit need not be the limit of the minimizers

Convergence at every fixed θ says nothing about $\widehat{\theta}_m^*$

All parameters at once

■ UNIFORM LAW OF LARGE NUMBERS

$$\sup_{\theta \in \Theta} \left| \widehat{L}_m(\theta) - L(\theta) \right| \longrightarrow 0 \quad \text{almost surely}$$

★ Θ compact, $\ell(\xi, \cdot)$ continuous, and dominated by an integrable envelope

■ CONSEQUENCE

$$\widehat{L}_m^* \longrightarrow L^* \quad \text{and} \quad \widehat{\theta}_m^* \longrightarrow \theta^* \quad \text{almost surely}$$

Uniform convergence carries convergence of values into convergence of minimizers

The optimal value is biased downward

■ A MINIMUM IS BELOW ANY PARTICULAR VALUE

$$\hat{L}_m^* = \underset{\theta}{\text{minimize}} \hat{L}_m(\theta) \leq \hat{L}_m(\theta^*)$$

■ TAKE EXPECTATIONS

$$\mathbb{E} \left[\hat{L}_m^* \right] \leq \mathbb{E} \left[\hat{L}_m(\theta^*) \right] = L(\theta^*) = L^*$$

★ θ^* is fixed, so $\hat{L}_m(\theta^*)$ is an unbiased estimate of $L(\theta^*)$

$$\mathbb{E} \left[\hat{L}_m^* \right] \leq L^* \text{ -- the training error understates the test error}$$

The bias shrinks with the sample size

Theorem: $\mathbb{E}[\hat{L}_m^*] \leq \mathbb{E}[\hat{L}_{m+1}^*]$ for every $m \geq 1$

■ PROOF

★ $\hat{L}_m^{(-j)}$ – the average over the m samples other than ξ_j

$$\hat{L}_{m+1}(\theta) = \frac{1}{m+1} \sum_{j=1}^{m+1} \hat{L}_m^{(-j)}(\theta)$$

$$\implies \hat{L}_{m+1}^* = \underset{\theta}{\text{minimize}} \hat{L}_{m+1}(\theta) \geq \frac{1}{m+1} \sum_{j=1}^{m+1} \hat{L}_m^{(-j)*}$$

★ each of the $m+1$ minima has the law of $\hat{L}_m^*$ – now take expectations

The gap closes from below: $\mathbb{E}[\hat{L}_m^*] \uparrow L^*$

Rate of convergence: the expansion

Curvature: $H = \nabla^2 L(\theta^*) \succ 0$

Gradient noise: $\Sigma = \text{Cov}(\nabla \ell(\xi, \theta^*))$

■ EXPAND THE OPTIMALITY CONDITION AROUND θ^*

$$0 = \nabla \hat{L}_m(\hat{\theta}_m^*) = \nabla \hat{L}_m(\theta^*) + \nabla^2 \hat{L}_m(\bar{\theta}) (\hat{\theta}_m^* - \theta^*)$$

$$\implies \sqrt{m} (\hat{\theta}_m^* - \theta^*) = - \left[\nabla^2 \hat{L}_m(\bar{\theta}) \right]^{-1} \left(\sqrt{m} \nabla \hat{L}_m(\theta^*) \right)$$

★ $\bar{\theta}$ lies on the segment between $\hat{\theta}_m^*$ and θ^*

The error is a ratio: a central limit term over a law of large numbers term

Rate of convergence: the limit law

■ NUMERATOR: A CENTRAL LIMIT THEOREM

$$\mathbb{E} [\nabla \ell(\xi, \theta^*)] = \nabla L(\theta^*) = 0 \quad \implies \quad \sqrt{m} \nabla \hat{L}_m(\theta^*) \implies \mathcal{N}(0, \Sigma)$$

■ DENOMINATOR: A LAW OF LARGE NUMBERS

$$\nabla^2 \hat{L}_m(\bar{\theta}) \longrightarrow H \quad \text{almost surely}$$

■ SLUTSKY

$$\sqrt{m} \left(\hat{\theta}_m^* - \theta^* \right) \implies \mathcal{N} \left(0, H^{-1} \Sigma H^{-1} \right)$$

The parameter error has scale $1/\sqrt{m}$

The excess risk is one order better

■ EXPAND THE EXPECTED RISK

$$L(\hat{\theta}_m^*) - L(\theta^*) \approx \frac{1}{2} (\hat{\theta}_m^* - \theta^*)^\top H (\hat{\theta}_m^* - \theta^*)$$

★ $\nabla L(\theta^*) = 0$ – the first order term is absent

■ TAKE EXPECTATIONS

$$\mathbb{E} \left[L(\hat{\theta}_m^*) - L(\theta^*) \right] \approx \frac{\text{tr}(H^{-1}\Sigma)}{2m}$$

Parameters converge at $1/\sqrt{m}$, while the risk at $1/m$

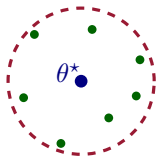
Another reason to regularize

■ THE CONSTANT IS $H^{-1}\Sigma H^{-1}$

- ★ a flat direction of L is a small eigenvalue of H , and a large variance

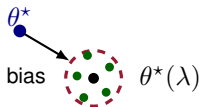
■ RIDGE RAISES EVERY EIGENVALUE BY λ

$$L_{\lambda}(\theta) = L(\theta) + \frac{\lambda}{2}\|\theta\|^2 \quad \implies \quad \nabla^2 L_{\lambda} = H + \lambda I$$



$\lambda = 0$

no bias, large variance



$\lambda > 0$

bias, small variance

Regularization trades bias for variance, and λ is set by cross validation

Where sample average approximation runs out

Practice: $\underset{\theta}{\text{minimize}} \quad \widehat{L}_m(\theta) = \frac{1}{m} \sum_{i=1}^m \ell(\xi_i, \theta)$

$$\theta_{k+1} = \theta_k - \alpha_k \frac{1}{m} \sum_{i=1}^m \nabla \ell(\xi_i, \theta_k)$$

■ EVERY STEP USES THE WHOLE SAMPLE

- ★ one gradient of $\widehat{L}_m$ costs m gradients of ℓ

■ THE SAMPLE IS FIXED IN ADVANCE

- ★ samples that arrive over time change the problem being solved

■ THE SAMPLE MUST BE STORED

- ★ m can exceed the memory available

QUESTION: Can each sample be used once, and never stored?

Questions?

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