Large Scale Optimization for Machine Learning

Meisam Razaviyayn
Lecture 12
razaviya@usc.edu
Agenda

• Review

• Regularization

• Cross validation
  • Parameter tuning
  • Termination criteria of optimization algorithms

• Structure of ERM
Recap: Empirical Risk Minimization

Predicting an output $y \in \mathcal{Y}$ given an input $x \in \mathcal{X}$, e.g., $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} \in \{0, 1\}$

Set of hypotheses: $\mathcal{H}$ with $h \in \mathcal{H}$ maps $\mathcal{X}$ to $\mathcal{Y}$

Loss function: $\ell : (\mathcal{X} \times \mathcal{Y}) \times \mathcal{H} \mapsto \mathbb{R}$

Data generating distribution $\mathbb{P}^*$ with $(x, y) \sim \mathbb{P}^*$

**Expected risk/Test error:** $L(h) \triangleq \mathbb{E}_{\mathbb{P}^*} [\ell((x, y), h)]$ \hspace{2cm} $h^* \in \arg \min_{h \in \mathcal{H}} L(h)$ \hspace{1cm} *Best we can hope for*

Training samples: $(x_1, y_1), \ldots, (x_n, y_n)$

**Empirical risk/Training error:** $\hat{L}(h) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), h)$ \hspace{2cm} $\hat{h} \in \arg \min_{h \in \mathcal{H}} \hat{L}(h)$ \hspace{1cm} *Empirical Risk Minimizer*

**Expected risk of ERM:** $L(\hat{h})$
Recap: What Set of Hypotheses?

Trade-offs between “number of samples”, “Expected risk or ERM”, “Complexity of hypothesis class”

**Occam's razor** (William of Ockham)
Recap: What Set of Hypotheses?

There are different ways of measuring complexity of a hypothesis class, but in general this trade-off exists
Regularization

• Example: regression

\[ \mathcal{H} = \{ h(x) = w^T x \mid w \in \mathbb{R}^d \} \]

\[ \min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2 \]

\[ \mathcal{H} = \{ h(x) = w^T x \mid \|w\|_2^2 \leq \beta \} \]

\[ \min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2 \]

s.t. \[ \|w\|_2^2 \leq \beta \]

\[ \min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_2^2 \]
Regularization

- **Goal**: reducing generalization error (expected risk) by reducing the complexity of $\mathcal{H}$

Empirical Risk Minimization

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), h)$$

Regularized Empirical Risk Minimization

$$\min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), h) + \lambda \mathcal{R}(h)$$

- Examples: Tikhonov regularization, total variation (TV) regularization, ...

- What happens if we heavily regularize?

- How to
  - Find the correct regularizer?
  - Find the correct hypotheses class $\mathcal{H}$?
Cross Validation

• **Goal**: avoiding over/under fitting

• **Strategy**: leave some of your data for the evaluation of your fitted model $\hat{h}$

**Popular Cross Validation Strategies:**

• K-fold cross validation
  • Partition the samples to K partitions
  • Use K-1 partitions for training and one for validation
  • Repeat K times over all partitioning and take the average

• Leave-m-out
  • Choose m out of n samples and use them for validation and the rest for training
  • Repeat over all $C(n,m)$ partitions and take the average (or randomly select)
  • Case m=1 is equivalent to n-fold cross validation → Leave-one-out

• Slightly biased, but still very helpful!
Cross validation can be used for:

• Choosing model fitting strategy
  • Example: SVM or logistic regression

• Type of regularization
  • Example: $L_2$ or $L_1$ norm

• Weight of the regularizer

• Stopping criteria

• Many other examples
Remarks

Regularization reduces model complexity

How to estimate expected risk and select models? Cross Validation!

Structure of ERM$	ext{s}$

- Summation/expectations in the objective
  - Stochastic optimization
  - Online optimization
  - Incremental methods

- Large number of blocks
  - Block optimization methods

\[
\min_w \sum_{i=1}^{n} (w^T x_i - y_i)^2
\]

\[
\min_w \sum_{i=1}^{n} \left( \log \left(1 + \exp(w^T x_i)\right) - y_i w^T x_i \right)
\]

\[
\min_{w,v} \frac{1}{n} \sum_{i=1}^{n} \max \left\{ 0, 1 - y_i (w^T x_i + v) \right\} + \lambda \|w\|_2^2
\]
Stochastic Optimization Framework

\[ h^* = \arg \min_{h \in \mathcal{H}} \mathbb{E}_{(x,y) \sim \mathcal{P}^*} [\ell((x, y), h)] \quad \hat{h}_n = \arg \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), h) \]

\[ w^* = \arg \min_w \mathbb{E}_\xi [\ell(\xi, w)] \quad \hat{w}_n = \arg \min_w \frac{1}{n} \sum_{i=1}^{n} \ell(\xi_i, w) \]

**Different names:** Empirical Risk Minimization, Sample Average Approximation

**Assumption:** Uniqueness of minimizer

**What is the relation between the optimal \( w^* \) and estimated \( \hat{w}_n \)?**
Sample Average Approximation (SAA)

\[ w^* = \arg \min_w \mathbb{E}_\xi [\ell(\xi, w)] \]

\[ \hat{w}_n = \arg \min_w \frac{1}{n} \sum_{i=1}^{n} \ell(\xi_i, w) \]

- For any fixed \( w \), Law of large number implies

\[ \hat{L}_n(w) \to L(w) \text{ as } n \to \infty \text{ almost surely} \]

- Under some regularity conditions, by uniform convergence of LLN:

\[ \hat{w}_n \to w^* \text{ as } n \to \infty \text{ almost surely} \]
Sample Average Approximation (SAA)

\[
\begin{align*}
  w^* &= \arg \min_w E_\xi [\ell(\xi, w)] \\
  L(w) &= \frac{1}{n} \sum_{i=1}^{n} \ell(\xi_i, w) \\

  \nu^* &= \min_w L(w) \\
  \nu_n &= \min_w \hat{L}_n(w)
\end{align*}
\]

- Under some regularity conditions, by LLN: \( \nu_n \rightarrow \nu^* \)

**Theorem:** For all \( n \geq 1 \), \( E[\nu_n] \leq E[\nu_{n+1}] \)

Training error is typically an under-estimator of the test error

Proof?
SAA: Rate of Convergence

\[ w^* = \arg \min_w L(w) \]
\[ \hat{w}_n = \arg \min_w \frac{1}{n} \sum_{i=1}^{n} \ell(\xi_i, w) \]

Assume a strongly convex quadratic objective

\[ \hat{w}_n = \arg \min_w \hat{L}_n(w^*) + (w - w^*)^T \nabla \hat{L}_n(w^*) + \frac{1}{2}(w - w^*)^T \nabla^2 \hat{L}_n(w^*)(w - w^*) \]

\[ \Rightarrow \sqrt{n}(\hat{w}_n - w^*) = - \left( \nabla^2 \hat{L}_n(w^*) \right)^{-1} \left( \sqrt{n} \nabla \hat{L}_n(w^*) \right) \]

\[ \Rightarrow \sqrt{n}(\hat{w}_n - w^*) \rightarrow \mathcal{N}(0, H^{-1} \Sigma H^{-1}) \]

Slutsky’s theorem

Want Hessian with large eigenvalues \(\Rightarrow\) Another justification for regularization

Also true for general case (under some regularity condition)