Large Scale Optimization for Machine Learning

Meisam Razaviyayn
Lecture 13
razaviya@usc.edu
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: $L(h) \triangleq \mathbb{E}_{\mathbb{P}^*}[\ell((x, y), h)]$

Empirical risk/Training error: $\hat{L}(h) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell((x_i, y_i), h)$

Expected risk of ERM: $L(\hat{h})$

Best we can hope for

Empirical Risk Minimizer
SAA: Rate of Convergence

\[ w^* = \arg \min_w L(w) \]

Assume a strongly convex quadratic objective

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

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

Slutsky's theorem

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

\[ \sqrt{n} \nabla \hat{L}_n(w^*) \text{ converges to } \mathcal{N}(0, \Sigma) \text{ in probability} \]

\[ \nabla^2 \hat{L}_n(w^*) \text{ converges to } H \triangleq \nabla^2 L(w^*) \text{ a.s.} \]

\[ \hat{w}_n - w^* = O_p\left(1/\sqrt{n}\right) \]

Also true for general case (under some regularity condition)
SAA: Rate of Convergence

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

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

Assume a strongly convex objective

\[ \hat{w}_n - w^* = O_p(1/\sqrt{n}) \]

\[ \sqrt{n}(\hat{w}_n - w^*) \to \mathcal{N}(0, \nabla^2 L(w^*)^{-1} \text{Cov} [\nabla \ell(\xi, w^*)] \nabla^2 L(w^*)^{-1}) \]

Is this order tight?

Want Hessian with large eigenvalues \( \rightarrow \) Another justification for regularization
Role of the Regularizer

\[
\begin{align*}
\mathbf{w}^* &= \arg \min_{\mathbf{w}} \mathbb{E}_\xi [\ell(\xi, \mathbf{w})] \\
\hat{\mathbf{w}}_n &= \arg \min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(\xi_i, \mathbf{w})
\end{align*}
\]

\[
\begin{align*}
\mathbf{w}^*(\mathcal{R}) &= \arg \min_{\mathbf{w}} \mathbb{E} \left( \ell(\xi, \mathbf{w}) + \lambda \|\mathbf{w}\|^2_2 \right) \\
\hat{\mathbf{w}}_n(\mathcal{R}) &= \arg \min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(\xi_i, \mathbf{w}) + \lambda \|\mathbf{w}\|^2_2
\end{align*}
\]

\[\lambda = 0\]

Over-fitting

\[\lambda \text{ large}\]

Under-fitting
SAA: Drawbacks

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

- What happens if we observe samples over time?
- What happens if we don’t have enough memory to store all samples?
Stochastic Approximation

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

- What if we don’t wait for \( n \) samples?
- Online gradient descent/Stochastic Approximation:

\[ w^{r+1} = w^r - \alpha^r \nabla \ell(\xi_r, w^r) \]

- Does it work with constant step-size rule?
- There is hope for diminishing step-size rule
- Can we use it in finite sample case? Limited memory scenario?
Stochastic Approximation: Analysis

\[ w^{r+1} = w^r - \alpha^r \nabla \ell(\xi_r, w^r) \quad w^* = \arg\min_w \mathbb{E}_\xi[\ell(\xi, w)] \]

- Assume:
  
  Strongly convex \( L(w) \)
  \[ \mathbb{E}_\xi[\|\nabla \ell(\xi, w)\|_2^2] \leq M^2, \quad \forall w \]

- Analysis

\[
\frac{1}{2}\|w^{r+1} - w^*\|^2 = \frac{1}{2}\|w^r - w^*\|^2 - \alpha^r (w^r - w^*)^T \nabla \ell(\xi_r, w^r) + \frac{1}{2}(\alpha^r)^2 \|\nabla \ell(\xi_r, w^r)\|^2
\]

Define \( \gamma^r = \frac{1}{2} \mathbb{E}[\|w^r - w^*\|^2] \)

\[
\Rightarrow \gamma^{r+1} \leq \gamma^r - \alpha^r \mathbb{E}[(w^r - w^*)^T \nabla \ell(\xi_r, w^r)] + \frac{1}{2}(\alpha^r)^2 M^2
\]
Stochastic Approximation: Analysis

\[ w^{r+1} = w^r - \alpha^r \nabla \ell(\xi_r, w^r) \]

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

\[ \Rightarrow \gamma^{r+1} \leq \gamma^r - \alpha^r \mathbb{E} [(w^r - w^*)^T \nabla \ell(\xi_r, w^r)] + \frac{1}{2} (\alpha^r)^2 M^2 \]

\[ \mathbb{E} [(w^r - w^*)^T \nabla \ell(\xi_r, w^r)] = \mathbb{E}_{\xi_{1:r-1}} [\mathbb{E}_{\xi_r} [(w^r - w^*)^T \nabla \ell(\xi_r, w^r) \mid \xi_{1:r-1}]] \]

\[ = \mathbb{E}_{\xi_{1:r-1}} [(w^r - w^*)^T \nabla L(w^r)] \]

\[ \geq \mathbb{E} [L(w^r) - L(w^*) + \frac{\sigma}{2} \|w^r - w^*\|^2] \]

\[ \geq \mathbb{E} [\sigma \|w^r - w^*\|^2] = 2\sigma \gamma^r \]

\[ \Rightarrow \gamma^{r+1} \leq (1 - 2\sigma\alpha^r) \gamma^r + \frac{1}{2} (\alpha^r)^2 M^2 \]

Setting \( \alpha^r = \frac{1}{r\sigma} \) and a neat induction implies

Better/Worse than before? \( \gamma^r \leq \frac{\theta}{2r} \), where \( \theta \triangleq \max \{ \|w^1 - w^*\|^2, \frac{M^2}{\sigma^2} \} \)
Stochastic Approximation: Analysis

\[ w^{r+1} = w^r - \alpha^r \nabla \ell(\xi_r, w^r) \]

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

\[
\mathbb{E}[\|w^r - w^*\|^2] \leq \frac{\theta}{2r}, \text{ where } \theta \triangleq \max\{\|w^1 - w^*\|, \frac{M^2}{\sigma^2}\}
\]

- Can be viewed as a noisy gradient descent
- Memory requirement

**Drawbacks:**

- Requires strong convexity
- Requires knowledge of strong convexity modulus
Diminishing step-size rule could be very sensitive

\[ \min_x \frac{1}{2}cx^2 \quad \text{with } c = 0.2 \]

\[ x^1 > 0 \text{ and } \alpha^r = \frac{1}{r} \Rightarrow x^{r+1} = x^1 \prod_{i=1}^{r} \left(1 - \frac{c}{i}\right) = O\left(r^{-1/5}\right) \]

Even after 10,000,000,000 iterations still not really close

\[ x^1 > 0 \text{ and optimal choice } \alpha^r = \frac{c^{-1}}{r} \Rightarrow \text{converges in one iteration.} \]

Diminishing is sensitive \(\Rightarrow\) If we use constant step-size, how far we are from optimality?