Summary: Optimization Methods for Large-Scale Machine Learning

V. Estellers

WS 2017

Statistical Learning Problem

Given a set of examples $(x_1, y_1), \ldots, (x_n, y_n)$

- each example $\xi = (x, y)$ is a pair of an input x and a scalar output y.
- loss $\ell(\hat{y}, y)$ measures the cost of predicting $\hat{y}$ when the answer is y
- family $\mathcal{H}$ of functions $h(\cdot; w)$ parametrized by a weight vector w.

We seek $h \in \mathcal{H}$ that minimizes the loss $f(\xi; w) = \ell(h(x; w), y)$.
Although we would like to average over the unknown distribution $P(x, y)$

$$f(w) = R(w) = \mathbb{E}[\ell(h(x; w), y)] = \int \ell(h(x; w), y) \mathrm{d}P(x, y)$$

we must settle for computing the average over the samples

$$f(w) = R_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i; w), y_i).$$

Statistical learning theory (Vapnik and Chervonenkis, 1971) justifies minimizing R_n instead of R when $\mathcal{H}$ is sufficiently restrictive.

Stochastic Gradient Method

The objective function $F\colon \mathbb{R}^d \mapsto \mathbb{R}$ can be the expected or empirical risk:

$$F(w) = \mathbb{E}[f(w,\xi)] \qquad \text{or} \qquad F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w).$$

The analysis applies to both objectives, depending on how the stochastic gradient estimates are chosen.

Stochastic Gradient Method

Choose an initial iterate w_1
for k=1,2,... **do**
  Generate a realization of the random variable ξ_k
  Compute a stochastic vector $g(w_k, \xi_k)$
  Choose a stepsize $\alpha_k > 0$
  Set the new iterate as $w_{k+1} = w_k - \alpha_k g(w_k, \xi_k)$
end for

Convergence of SG

Two fundamental lemmas bound the descent of stochastic iterations by deterministic values.

Theorem

If F is L-smooth and c-strongly convex and there are $M \leq 0$ and $M_G \geq \mu^2 \geq 0$ such that $\mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|^2] \leq M + M_G \|\nabla F(w_k)\|^2$, then

- *SG with a positive stepsize $\alpha \leq \frac{\mu}{LM_G}$ converges linearly to a $\frac{\alpha LM}{2c\mu}$-neighborhood of a minimizer of F with rate $\mathcal{O}\left((1 - \alpha c\mu)^{k-1}\right)$.*

- *if the stepsizes decrease as $\alpha_k = \frac{\beta}{\gamma + k}$ for some $\beta > \frac{1}{c\mu}$, $\gamma > 0$ then SG converges to a minimizer of F with rate $\mathcal{O}\left(\frac{1}{k}\right)$.*

Stochastic vs Batch Methods

Let R_n^* be the minimal value of R_n, then if R_n is strongly convex

- the error of a batch gradient method satisfies

$$|R_n(w_k) - R_n^*| \leq \mathcal{O}(\rho^k), \quad \rho \in (0,1).$$

  The number of iterations where the training error is above ϵ is proportional to $\log(\frac{1}{\epsilon})$, and the cost of ϵ-optimality is $\mathcal{O}(n \log(\frac{1}{\epsilon}))$.

- the SG error for i_k is drawn uniformly from $\{1, \ldots, n\}$ is

$$\mathbb{E}[|Rn(w_k) - R_n^*|] = \mathcal{O}(\frac{1}{k}) \tag{0.1}$$

  As it does not depend on n, the cost of ϵ-optimality is $\mathcal{O}(\frac{1}{\epsilon})$.

The SG cost $\mathcal{O}(\frac{1}{\epsilon})$ is smaller than the batch cost $\mathcal{O}(n \log(\frac{1}{\epsilon}))$ if n is large.

Stochastic vs Batch Methods

SG avoids overfitting in the sense that the minimizer of the empirical risk found by SG has some minimization guarantees on the expected risk.

By applying the SG iteration with $\nabla f(w_k; x_{i_k})$ replaced by $\nabla f(w_k; \xi_k)$ with each ξ_k drawn independently according to the distribution P,

$$\mathbb{E}[|R(w_k) - R^*|] = \mathcal{O}(\frac{1}{k}). \tag{0.2}$$

This is again a sublinear rate, but on the expected risk.

Noise-Reduction Methods

Noise-Reduction Methods: instead of decreasing the learning rate to converge to the optimum, reduce variance of the stochastic gradients. They achieve a linear convergence rate at a higher per-iteration cost.

Other methods come with few guarantees but work well in practice:

- Gradient Methods with Momentum
- Accelerated Gradient Method
- Adaptive Methods: adagrad, adadelta, adam