Pratical issues: optimization and learning

Sham M Kakade
© 2018

University of Washington
cse446-staff@cs.washington.edu
Announcements

▶ Midterm summary:
  ▶ stats: 71.5 std: 18
  ▶ Office hours today: 1:15-2:30 (No office hours on Monday)

▶ Monday: John Thickstun guest lecture

▶ Grading:
▶ HW3 posted
  ▶ will be periodically updated for typos/clarifications
  ▶ extra credit posted soon

▶ Today:
  ▶ Midterm review
  ▶ GD/SGD: practical issues
Midterm
What is a good model of this distribution?
What is a good model of this distribution?
“A mixture of Gaussians”
Midterm Q4: scratch space
Midterm: scratch space
Midterm Q5: scratch space
Midterm: scratch space
Today
The “general” Loss Minimization Problem

\[ \mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \ell(x_n, y_n, \mathbf{w}) + R(\mathbf{w}) \]

How do we run GD? SGD? Which one to use?

How do run them?
Our running example

$$\arg\min_w \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2} (y_n - w \cdot x_n)^2 + \frac{1}{2} \lambda \|w\|^2$$

- GD? SGD?

- Note we are computing an average. What is a crude way to estimate an average?

Will it converge?
How does GD behave? A 1-dim example
GD: How do we set the step sizes?

- Theory:
  - square loss:
  - more generally:

- Practice:
  - square loss:
  - more generally:

- Do we decay the stepsize?
SGD for the square loss

**Data:** step sizes $\langle \eta^{(1)}, \ldots, \eta^{(K)} \rangle$

**Result:** parameter $w$

initialize: $w^{(0)} = 0$;

for $k \in \{1, \ldots, K\}$ do

$n \sim \text{Uniform}\{\{1, \ldots, N\}\}$;

$w^{(k)} = w^{(k-1)} + \eta^{(k)} (y_n - w^{(k-1)} \cdot x_n) x_n$;

end

return $w^{(K)}$;

**Algorithm 1:** SGD
SGD for the square loss

Data: step sizes $\langle \eta^{(1)}, \ldots, \eta^{(K)} \rangle$

Result: parameter $w$

initialize: $w^{(0)} = 0$;

for $k \in \{1, \ldots, K\}$ do

\[
\begin{align*}
    n &\sim \text{Uniform}\{1, \ldots, N\}; \\
    w^{(k)} &\equiv w^{(k-1)} + \eta^{(k)} (y_n - w^{(k-1)} \cdot x_n) \cdot x_n;
\end{align*}
\]

end

return $w^{(K)}$;

Algorithm 2: SGD

- where did the $N$ go?
- regularization?
- minibatching?
SGD: How do we set the step sizes?

- Theory:
- Practice:
  - How do start it?
  - When do we decay it?
Stochastic Gradient Descent: Convergence

\[ w^* = \arg\min_w \frac{1}{N} \sum_{n=1}^N \ell_n(w) \]

- \( w^{(k)} \): our parameter after \( k \) updates.
- Thm: Suppose \( \ell(\cdot) \) is convex (and satisfies mild regularity conditions). There is a decreasing sequence of step sizes \( \eta^{(k)} \) so that our function value, \( F(w^{(k)}) \), converges to the minimal function value, \( F(w^*) \).
- GD vs SGD: we need to turn down our step sizes over time!
Making features: scratch space