# 5. Stochastic Gradient Descent

Exercises due Sep 30, 2020 05:29 IST *Completed*

# Stochastic Gradient Descent


Start of transcript. Skip to the end.

Now our problem looks like this.

We have an average of the training losses

plus regularization term.

And we can write it by moving the regularization term

inside the average, since it doesn't

**Video**
Download video file

**Transcripts**
Download SubRip (.srt) file
Download Text (.txt) file

## SGD and Hinge Loss

1/1 point (graded)
As we saw in the lecture above,

$$J(\theta, \theta_0) = \frac{1}{n}\sum_{i=1}^{n} \text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) + \frac{\lambda}{2}\,||\,\theta\,||^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) + \frac{\lambda}{2}\,||\,\theta\,||^2\right]$$

With stochastic gradient descent, we choose $i \in \{1, \ldots, n\}$ at random and update $\theta$ such that

$$\theta \leftarrow \theta - \eta \nabla_\theta \left[\text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) + \frac{\lambda}{2}\,||\,\theta\,||^2\right]$$

What is $\nabla_\theta \left[\text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right)\right]$ if $\text{Loss}_h\left(y^{(i)}\left(\theta \cdot x^{(i)} + \theta_0\right)\right) > 0$?

- ◯ $y^{(i)}x^{(i)}$
- ◉ $-y^{(i)}x^{(i)}$
- ◯ $0$
- ◯ $\lambda\theta$
- ◯ $-\lambda\theta$

✔

Submit You have used 1 of 3 attempts

## Comparison with Perceptron

1/1 point (graded)
Observing the update step of SGD,

$$\theta \leftarrow \theta - \eta \nabla_\theta \left[ \text{Loss}_h \left( y^{(i)} \left( \theta \cdot x^{(i)} + \theta_0 \right) \right) + \frac{\lambda}{2} || \theta ||^2 \right]$$

Which of the following is true?

- As in perceptron, $\theta$ is not updated when there is no mistake
- Differently from perceptron, $\theta$ is updated even when there is no mistake ✔

Submit You have used 1 of 1 attempt

## Discussion

Hide Discussion

**Topic:** Unit 1 Linear Classifiers and Generalizations (2 weeks):Lecture 4. Linear Classification and Generalization / 5. Stochastic Gradient Descent

Add a Post

Show all posts by recent activity

| | | |
|---|---|---|
| ? | How much more efficient can SGD get compared with normal GD?<br>I am thinking although SGD calculates gradient faster, but I guess it will take more iterations to converge, how much more efficient? I... | 3 |
| ? | How doe this reduce the Variance?<br>First of all, what is exactly meant by the "variance due to stochasticity" and why does having a square summable sequence of learnin... | 6 |
| ? | when do we see solution<br>when is the solution with explanation posted? | 2 |
| ? | How do you calculate gradient descend if my hinge loss term at the non-differentiable point?<br>Hi, I may be missing something here. Hypothetically, if my randomly selected sample happens to be at the sharp point (z=1 agreeme... | 3 |
| 💬 | Regarding the omission of the offset parameter<br>Could anyone provide insight on whether and how these intuitions might change if the offset parameter were included? | 1 |
| ? | Learning rate reduction and its sum<br>In the lecture the professor mentioned that we have to decrease the learning rate while working with SGD, then said that what we h... | 3 |
| 💬 | Difference with perceptron<br>Professor mentions three difference with the perceptron algorithm: 1- using decreasing learning rate due to stochasticity, 2- answer ... | 1 |
| 💬 | is lambda * theta added to both the cases (when loss is 0 and loss is greater than 0)?<br>It looks from his writing that it is added only when loss is greater than 0 but my understanding is that he is adding it on both case. Is ... | 2 |
| ? | Can SGD never reach global maximum/minimum?<br>Because the SGD uses random samples (or shuffle sample order), does that mean that using SGD we may never reach the true opti... | 2 |
| ? | Can "SGD" be used for other objective functions?<br>Does "SGD" refer to the entire algorithm, comprising both the cost function and the update law? Generally when we talk about "grad... | 3 |
| ? | Clarification about the 3rd difference b/w SGD and Perceptron<br>Professor lists the first two differences b/w SGD and Perceptron. [Not stated here to avoid giving answers] Third difference: "And the... | 3 |