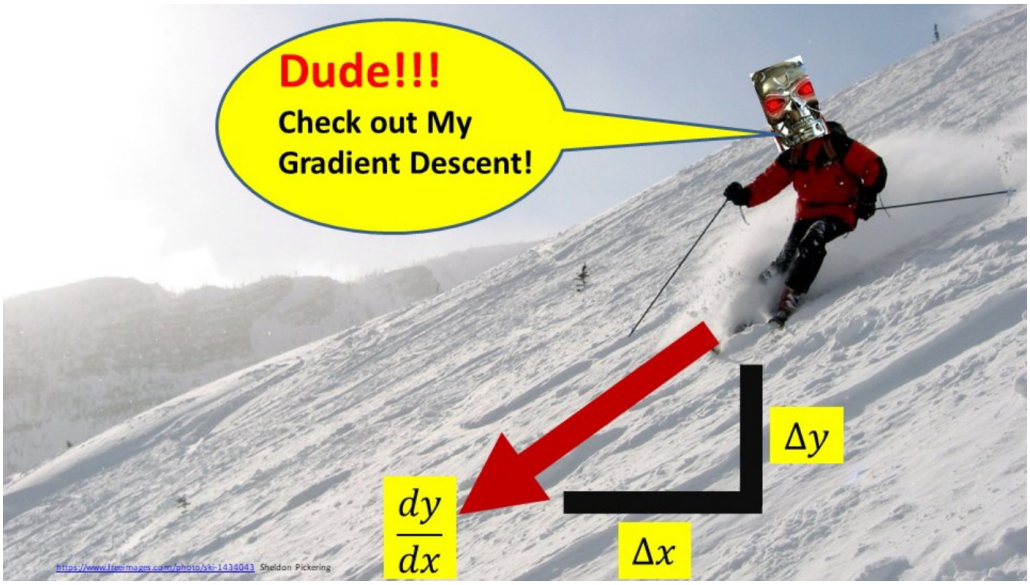
Notes: Stochastic Gradient Descent II



1 Pre-lecture Work

Problem 1. I recommend watching the following StatQuest videos if you are unfamiliar with either gradient descent or stochastic gradient descent.

1. Gradient descent: `https://www.youtube.com/watch?v=sDv4f4s2SB8`
2. Stochastic gradient descent: `https://www.youtube.com/watch?v=vMh0zPT0tLI`

The following StatQuest videos about regularization may also be helpful:

1. L2 regularization: `https://www.youtube.com/watch?v=Q81RR3yKn30`
2. L1 regularization: `https://www.youtube.com/watch?v=NGf0voTMlcs`
3. Elastic-net regularization: `https://www.youtube.com/watch?v=1dKRdX9bfIo`

Problem 2. Before the *2nd day of lecture*, you should complete all the problems marked with a $\star$.

2 Lecture Warm-up

Problem 3. Recall that in a regression problem, feature vectors are contained in the space $\mathcal{X} = \mathbb{R}^d$, and the prediction space is $\mathcal{Y} = \mathbb{R}$. In Tikhonov regression, we use the squared loss and the L2 regularizer. The resulting minimization problem is

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2}(\mathbf{w}^T \mathbf{x}_i - y)^2 + \lambda \|\mathbf{w}\|_2^2. \tag{1}$$

1. Find a closed form solution for $\hat{\mathbf{w}}$.

2. What is the runtime of computing the above expression?

3. Why is there a closed form solution for $\hat{\mathbf{w}}$?

3 Lecture

Problem 4. Convex-Lipschitz Gradient Descent.

1. The update rule (Eq. 14.1)

2. Corollary 14.2 (convergence rate)

3. How should we choose ϵ and T?

Problem 5. Lipschitz Stochastic Gradient Descent.

1. The update rule (Section 14.3)

2. ($\star$) Definition 12.12 (convex-lipschitz-bounded learning problem)

3. ($\star$) Theorem 14.8 (optimization convergence rate)

4. ($\star$) Corollary 14.12 (learning convergence rate)

Problem 6. Convex Smooth Stochastic Gradient Descent.

1. The update rule is the same as in the Lipschitz case. The only difference is the analysis.
2. ($\star$) Definition 12.13 (convex-smooth-bounded learning problem)

3. ($\star$) Corollary 14.14 (learning convergence rate)

Problem 7. Strongly Convex Stochastic Gradient Descent.

1. The update rule is the same as in the Lipschitz case. The only difference is the analysis. In this case, the book presents the analysis in just a single theorem.
2. ($\star$) Theorem 14.11 (optimization convergence rate)

4 Learning Problems

Problem 8. Consider the classification problem using the logistic loss with no regularization.

1. Assume the parameters are restricted so that for all $\mathbf{w}$, $\|\mathbf{w}\|_2 \leq \sqrt{d}$; and all data points $\mathbf{x}_i$ satisfy $\|\mathbf{x}_i\|_2 \leq \sqrt{d}$. What is the generalization bound of SGD?

2. Assume that all data points satisfy $\|\mathbf{x}_i\|_2 \leq 1$. What is the generalization bound of SGD?

   Why is this significant?

3. Now assume that L2 regularization is used. What is the new generalization bound of SGD?

4. Now assume that L1 regularization is used. What is the new generalization bound of SGD?

Problem 9. Consider the regression problem using the squared loss with no regularization.

1. Assume the parameters are restricted so that for all $\mathbf{w}$, $\|\mathbf{w}\|_2 \leq \sqrt{d}$; and all data points $\mathbf{x}_i$ satisfy $\|\mathbf{x}_i\|_2 \leq \sqrt{d}$. What is the generalization bound of SGD?

2. Assume that all data points satisfy $\|\mathbf{x}_i\|_2 \leq 1$. What is the generalization bound of SGD?

   Why is this significant?

3. Now assume that L2 regularization is used. (This is the Tikhonov regularization problem from the warmup.) What is the new generalization bound of SGD?

4. Now assume that L1 regularization is used. (This is often called Lasso regression.) What is the new generalization bound of SGD?

Problem 10. (optional) Repeat the above exercises using the absolute loss and Huber loss for regression problems, and the hinge loss and exponential loss for classification problems. Also consider elastic net regularization for all combinations of losses.

Problem 11. This problem concerns generalization bounds for nearest neighbor algorithms.

1. ($\star$) Reproduce Theorem 19.3 (generalization bound of nearest neighbor)

2. How does this result relate to the VC-dimension of nearest neighbor being infinite?