Homework 1

Deep Learning 2025 Spring

Due: 11:59pm, 2025/3/10

1 True or False

Problem 1. The greatest advantage of residual connection is that it prevents overfitting.

Problem 2. Dropout and batch normalization can be used together.

Problem 3. BatchNorm and LayerNorm are both special cases of GroupNorm.

2 Q&A

Problem 4. (Descent Lemma) Given that the gradient of function $f : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous: for any $x, y \in \mathbb{R}^d$, there exists an $L(L > 0)$, such that $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$.
Prove descent lemma: $f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|x - y\|^2$.

Problem 5. (Convergence of Gradient Descent) Assume function $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, L-smooth and μ-strongly-convex. We apply gradient descent algorithm to find $x^\star = \arg\min_x f(x)$ starting from x^0, and $\|x^0 - x^\star\| = R$. Prove that to find a x^k such that $\|x^k - x^\star\| \leq \epsilon$ with learning rate $\eta = \frac{1}{L}$, the number of iterations (i.e., k) should have order $\mathcal{O}(\frac{L}{\mu}\log\frac{R}{\epsilon})$. The gradient descent algorithm is given by $x^{k+1} \leftarrow x^k - \eta\nabla f(x^k)$.

Problem 6. (Gradient Descent with Momentum) $f(x)$ is defined by

$$f(x) = \begin{cases} \frac{25}{2}x^2 & \text{if} \quad x < 1 \\ \frac{1}{2}x^2 + 24x - 12 & \text{if} \quad 1 \leq x < 2 \\ \frac{25}{2}x^2 - 24x + 36 & \text{otherwise.} \end{cases}$$

Given $\beta = \frac{4}{9}$, $\eta = \frac{1}{9}$, and initial value $x^0 = 3.3$, show that gradient descent with momentum does not converge in this case. Gradient descent with momentum is given by $x^{k+1} \leftarrow x^k - \eta\nabla f(x^k) + \beta(x^k - x^{k-1})$.

Problem 7. (Quadratic Convergence) Consider an iterative algorithm with quadratic convergence for minimizing a smooth, strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$. Suppose the sequence of iterates $\{x_k\}$ satisfies the quadratic convergence condition:

$$\|x_{k+1} - x^*\| \leq C \|x_k - x^*\|^2$$

where x^* is the optimal solution, $0 < C < 1$ is a constant, and the initial error is bounded by $\|x_0 - x^*\| \leq \delta$. Derive the number of iterations k required to guarantee that the error satisfies $\|x_k - x^*\| \leq \varepsilon$, where $\varepsilon > 0$ is a predefined tolerance.

Problem 8. (Second Order Optimization) Assume f is twice differentiable, L-smooth, and μ-strongly-convex. $\nabla^2 f$ is Lipchitz continuous. Consider $x^\star$ where $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star)$ is positive definite. Suppose x^0 is sufficiently close to $x^\star$. Prove that Newton's method converges to $x^\star$, and the convergence rate is quadratic. The Newton's method is given by $x^{k+1} \leftarrow x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$.

Problem 9. (Kaiming Initialization) Consider a neural network with linear layers $Z^l = W^l X^l$ and ReLU activation $X^l = \max(0, Z^{l-1})$. We make the following assumptions for every $l \geq 2$:

1. W^l and X^l are mutually independent.
2. The elements of W^l are i.i.d Gaussian variables with zero mean.
3. The elements of X^l are i.i.d.
4. $\mathbb{E}\left[Z_{ij}^l\right] = 0$.

Consider the *forward pass* of such a neural network. If we want the variance of hidden neurons to be unchanged from layer to layer, prove that the variance of W^l should be $\frac{2}{h^l}$ $(l \geq 2)$, where h^l is the number of neurons in the l-th hidden layer.