CS60021: Scalable Data Mining
Practice Problems: Optimization

1. Consider the following statements about SGD are true:

   (i) When mini-batch size become equal dataset size, SGD becomes a descent method.

   (ii) Addition of momentum term is expected to reduce number of updates.

   (a) True for both (i) and (ii).

   (b) True for (i) but false for (ii).

   (c) True for (ii) but false for (i).

   (d) False for both (i) and (ii).

2. SGD update for parameter w for regularized linear regression, given by the loss function below, is:

$$L(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \lambda \|w\|^2$$

   (a) * $w^{t+1} = w^t - \eta_t(-(y_i - w^T x_i).x_i + \frac{\lambda}{n} w^t)$

   (b) $w^{t+1} = w^t - \eta_t(-(y_i - w^T x_i).x_i + \lambda w^t)$

   (c) $w^{t+1} = w^t + \eta_t(y_i - w^T x_i).x_i$

   (d) $w^{t+1} = w^t + \eta_t(y_i - w^T x_i).x_i$ followed by a once per epoch update of $(\eta_t \lambda w)$

3. Which of the following learning rate schedules are guaranteed to result in convergent updates in the expected sense:

   (a) $\eta_t = \frac{1}{t}$

   (b) $\eta_t = \frac{1}{t^2}$

   (c) $\eta_t = \frac{1}{t^{0.5+\epsilon}}$

   (d) $\eta_t = \frac{1}{t^{0.5-\epsilon}}$

4. Which of the following non-differentiable functions have sub-gradients:

   (a) $f(x) = |x| + x$

   (b) $f(x) = |x|$

   (c) $f(x) = |x - t_1|$

   (d) $f(x) = |x - t_2| + |x - t_2|$

5. Which of the following statements is true:

(i) For the consensus optimization using ADMM, the updates to the consensus variable can be eliminated.

(ii) ADMM loss minimization can also be formulated as a fully distributed optimization over arbitrary graph topology.

(a) True for both (i) and (ii).

(b) True for (i) but false for (ii).

(c) True for (ii) but false for (i).

(d) False for both (i) and (ii).

6. Which of the following options are true:

(i) ADMM is a descent method.

(ii) ADMM is same as method of multipliers applied to multiple variables.

(a) Both (i) and (ii) are true

(b) (i) is true and (ii) is false

(c) (ii) is true and (i) is false

(d) Both (i) and (ii) are false

7. Which of the following options are true:

(i) Dual ascent converges only for strictly convex functions.

(ii) Augmented Lagrangian makes the objective function strictly convex.

(a) Both (i) and (ii) are true

(b) (i) is true and (ii) is false

(c) (ii) is true and (i) is false

(d) Both (i) and (ii) are false

8. Which of the following is the correct formula for augmented lagrangian for following problem:

$$\begin{aligned} &\min_{x,y} \;\; x^2 + y^2 \\ &\text{sub. to.} \;\; x = y \end{aligned}$$

(a) $L(x, y, \lambda, \rho) = x^2 + y^2 + \lambda(x - y) + \rho(x - y)^2$

(b) $L(x, y, \lambda, \rho) = x^2 + y^2 - \lambda(x - y) + \rho(x - y)^2$

(c) $L(x, y, \lambda, \rho) = x^2 + y^2 + \lambda(x + y) + \rho(x - y)^2$

(d) Both (a) and (b).

9. Which of the following ADMM update equations are wrong for the following problems:

$$\begin{aligned} &\min_{x,y} \;\; x^2 + y^2 \\ &\text{sub. to.} \;\; x = y \end{aligned}$$

(a) $x^{t+1} = \frac{\lambda^t + 2\rho y^t}{2(1+\rho)}$

(b) $y^{t+1} = \frac{-\lambda^t + 2\rho x^t}{1(1+\rho)}$

(c) $\lambda^{t+1} = \lambda^t + \rho(x^{t+1} - y^{t+1})$

(d) None of the above.

10. ADMM is a meta optimisation algorithm because:

    (a) It uses other optimizers to solve each step.

    (b) It can be used for distributed optimization.

    (c) It can be used for optimizing non-differentiable functions.

    (d) None of the above.

11. Which of the following about ADMM is true:

    (i) It can be used for distributed optimization.

    (ii) It can be used for optimizing non-differentiable functions.

    (a) Both (i) and (ii) are true

    (b) (i) is true and (ii) is false

    (c) (ii) is true and (i) is false

    (d) Both (i) and (ii) are false

12. Which of the following is true (in expectation) about minibatch SGD:

    (a) Increasing the minibatch size always leads to faster convergence in terms of wall clock time.

    (b) Increasing the minibatch size leads to lower variance in the norm of the averaged stochastic gradient, at any given time.

    (c) Increasing the minibatch size leads to lower fluctuations in terms of objective function value across updates.

    (d) None of the above.

13. For the problem of $\min_x \sum_{i=1}^{m} f_i(x)$, which of the following are valid gradient descent updates:

    (a) $x^{t+1} = x^t - \nabla_x f_i(x)$ for random $i \in \{1, ..., m\}$

    (b) $x^{t+1} = x^t - \frac{1}{t}\nabla_x f_i(x)$ for random $i \in \{1, ..., m\}$

    (c) $x^{t+1} = x^t - \frac{1}{\sqrt{t}}\nabla_x f_i(x)$ for random $i \in \{1, ..., m\}$

    (d) $x^{t+1} = x^t - \frac{1}{t^2}\nabla_x f_i(x)$ for random $i \in \{1, ..., m\}$

14. Which of the following is the Nesterov's accelerated gradient descent updates for the objective function $\min_x f(x) = \sum_{i=1}^{m} b_i(x - a_i)^2$:

    (a) $v^t = \frac{\eta}{(v^{t-1})^2}\sum_{i=1}^{m} 2b_i(x - a_i)$, followed by $x^{t+1} = x^t - v^t$

    (b) $v^t = \gamma v^{t-1} + \eta \sum_{i=1}^{m} 2b_i(x - a_i)$, followed by $x^{t+1} = x^t - v^t$

(c) $v^t = \gamma v^{t-1} + \eta \sum_{i=1}^{m} 2b_i(x - \gamma v^t - a_i)$, followed by $x^{t+1} = x^t - v^t$

(d) None of the above

15. Which of the following are true about the following function:

$$f(x, y) = (a - x)^2 + b(y - x^2)^2 \quad a, b \geq 0$$

(a) Convex in x but non-convex in y

(b) Convex in y but non-convex in x

(c) Convex in both x and y

(d) Non-convex in both x and y

16. Points on the curve $x = y^2 + 1$ which are closest to $(0, 2)$ are:

(a) One of the real solutions of $2y^3 + 3y - 2 = 0$

(b) One of the real solutions of $2y^3 - 3y - 2 = 0$

(c) One of the real solutions of $4y^3 - 2y + 1 = 0$

(d) One of the real solutions of $4y^3 + 2y + 1 = 0$

17. For the following optimization problem:

$$\min_{x_1, x_2} f(x_1) + g(x_2)$$
$$\text{s.t. } x_1 + x_2 = 0$$

which of the following lagrangians are used in the ADMM formulation:

(a) $L(x_1, x_2, \lambda) = f(x_1) + g(x_2) + \lambda(x_1 + x_2)$

(b) $L(x_1, x_2, \lambda) = f(x_1) + g(x_2) + \lambda(x_1 + x_2)$ s.t. $\lambda \geq 0$

(c) $L(x_1, x_2, \lambda) = f(x_1) + g(x_2) + \lambda(x_1 + x_2) + \rho \|x_1 + x_2\|^2$ where ρ is a parameter

(d) None of the above.

18. The dual decomposition updates for x_1, x_2 for the following problem, with λ as the largrange multiplier, are going to be:

$$\min_{x_1, x_2} f(x_1) + g(x_2)$$
$$\text{s.t. } x_1 + x_2 = 0$$

(a) $x_1^{t+1}, x_2^{t+1} = \arg\min_{x_1, x_2} f(x_1) + g(x_2) + \lambda^t(x_1 + x_2)$

(b) $x_1^{t+1} = \arg\min_{x_1} f(x_1) + \lambda^t x_1$ and $x_2^{t+1} = \arg\min_{x_2} f(x_2) + \lambda^t x_2$

(c) All of the above

(d) None of the above.

19. Which of the following statements about SGD and ADMM are true:

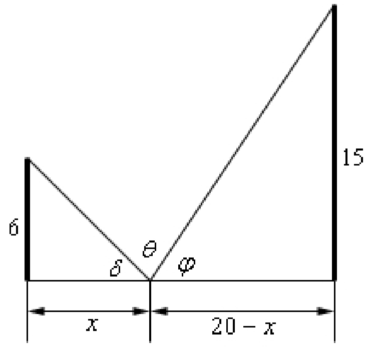
(a) Both ADMM and SGD are descent methods.

(b) ADMM is a descent method but SGD is not.

(c) SGD is a descent method but ADMM is not.

(d) Both ADMM and SGD are not descent methods.

20. Two poles, one 6 meters tall and one 15 meters tall, are 20 meters apart. A wire is attached to the top of each pole and it is also stacked to the ground somewhere between the two poles. Where should the wire be staked (x) so that the angle formed by the two pieces of wire at the stake (θ) is a maximum?



(a) $x = 9.8$

(b) $x = 10.8$

(c) $x = 8.8$

(d) $x = 0$

21. The following update rules for minimizing loss function $l(x)$, are from the method:

$$x_t = \gamma x_{t-1} + \eta \nabla_\theta l(\theta - \gamma x_{t-1})$$
$$\theta = \theta - x_t$$

(a) Momentum update.

(b) Nesterov acceleration.

(c) ADAM (Adptive moment estimation).

(d) RMSProp.

22. Which of the following are sub-gradients of hinge loss function $l(w) = max(0, yw^T x)$, where (x, y) are the true data-point:

(a) $\nabla_w l(w) = 0$ if $yw^T x < 0$ else yx.

(b) $\nabla_w l(w) = 0$ if $yw^T x \leq 0$ else yx.

(c) $\nabla_w l(w) = 0$.

(d) $\nabla_w l(w) = yx$.

23. Consider the following exponential loss function, which is sometimes used for binary classification: $l(w) = \frac{1}{n}\sum_{i=1}^{n} e^{-y_i w^T x_i}$. The mini-batch stochastic gradient descent update using mini-batch size of k are:

(a) $w^{t+1} = w^t + \eta_t \frac{1}{k} \sum_{i \in S_{n,k}} y_i x_i e^{-y_i w^T x_i}$ where $S_{n,k}$ is a k-sized random subset of $\{1, \ldots, n\}$

(b) $w^{t+1} = w^t + \eta_t \frac{1}{n} \sum_{i \in S_{n,k}} y_i x_i e^{-y_i w^T x_i}$ where $S_{n,k}$ is a k-sized random subset of $\{1, \ldots, n\}$

(c) $w^{t+1} = w^t + \eta_t \frac{k}{n} \sum_{i \in S_{n,k}} y_i x_i e^{-y_i w^T x_i}$ where $S_{n,k}$ is a k-sized random subset of $\{1, \ldots, n\}$

(d) $w^{t+1} = w^t + \eta_t \sum_{i \in S_{n,k}} y_i x_i e^{-y_i w^T x_i}$ where $S_{n,k}$ is a k-sized random subset of $\{1, \ldots, n\}$

24. Consider the following statements and answer the appropriate questions:

(i) ADMM uses the augmented Lagrange multiplier because that makes the sub-optimizations problems strictly convex.

(ii) Augmented Lagrangian of a separable loss function (expressed as a sum over component loss functions over variables) is a separable function.

(a) Both (i) and (ii) are true.

(b) (i) is true but (ii) is false.

(c) (ii) is true but (i) is false.

(d) Both (i) and (ii) are false.