Problem 1 (20 Points): Matrix Inverse Identities

Derive the following two useful identities involving matrix inverse:

1. **10 Points.** Suppose $\mathbf{Q} \in \mathbb{R}^{N\times N}$, $\mathbf{P} \in \mathbb{R}^{M\times M}$, and $\mathbf{B} \in \mathbb{R}^{M\times N}$, verifying the following identity:

$$(\mathbf{Q}^{-1} + \mathbf{B}^T\mathbf{P}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{P}^{-1} = \mathbf{Q}\mathbf{B}^T(\mathbf{B}\mathbf{Q}\mathbf{B}^T + \mathbf{P})^{-1}. \tag{1}$$

   Note: If $N \gg M$, it will be much cheaper to evaluate the right-hand side of Equation 1, which involves an inverse of a matrix $M \times M$, than the left-hand side, which involves an inverse of a matrix $N \times N$. A special case that is commonly used in machine learning is:

$$(\mathbf{I} + \mathbf{A}\mathbf{B})^{-1}\mathbf{A} = \mathbf{A}(\mathbf{I} + \mathbf{B}\mathbf{A})^{-1}. \tag{2}$$

2. **10 Points.** Verifying the following *Woodbury identity*:

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{3}$$

   Note: Equation 3 is useful when $\mathbf{A}$ is *large and diagonal* (hence easy to invert), and $\mathbf{B}$ (or $\mathbf{C}$) is a thin and tall (or fat and short) matrix.

Problem 2 (20 Points): Matrix Calculus

Let $\mathbf{x}$ and $\mathbf{y}$ be an n-dim vector and m-dim vector, respectively. Each entry y_i in $\mathbf{y}$ is a function of x_j in $\mathbf{x}$, or saying that $\mathbf{y}$ is a function $\mathbf{x}$. Then, we can define the derivative of $\mathbf{y}$ with respect to $\mathbf{x}$ and it is an $n \times m$ matrix as follows:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \tag{4}$$

1. **10 Points.** Given $\mathbf{x} = [x_1; x_2; x_3] \in \mathbb{R}^3$ and $\mathbf{y} = [y_1; y_2] \in \mathbb{R}^2$, where $y_1 = x_1^2 - x_2$ and $y_2 = x_3^2 + 3x_2$. Compute $\partial\mathbf{y}/\partial\mathbf{x}$.
2. **10 Points.** The transformation from spherical to Cartesian coordinates is defined by:

$$x = r\sin\theta\cos\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\theta, \tag{5}$$

   where $r > 0$, $0 < \theta < \pi$, and $0 \le \phi < 2\pi$.

   Let $\mathbf{x} = [x; y; z]$ and $\mathbf{y} = [r; \theta; \phi]$. Compute $\partial\mathbf{x}/\partial\mathbf{y}$.

Problem 3 (20 Points) Newton's Method for Solving Least Squares in Linear Regression

Prove that if we use Newton's method to compute the least square in linear regression, we only need one iteration to converge to the optimal parameters.

1. **10 Points.** Find the Hessian of the least square loss $L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i^T\mathbf{w} - y_i)^2$ with respect to parameter vector $\mathbf{w}$.
2. **10 Points.** Show that the first iteration of Newton's method gives us $\mathbf{w}^* = (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{y}$, the solution to our least squares problem.

Problem 4 (20 Points): Constrained Optimization vs. Unconstrained Optimization in Regularized Linear Regression

Use the techniques of Lagrangian multipliers (Please see Appendix E in CB).

1. **10 Points.** Showing that minimizing the ordinal least square loss subject to the l_p (l_2 or l_1) norm constraint is equivalent to minimizing the l_p (l_2 or l_1) regularized least square loss. In particular,

$$\min_{\mathbf{w}} L(\mathbf{w}) = \sum_{n=1}^{N} (f(\mathbf{x}_n; \mathbf{w}) - t_n)^2, \text{s.t.}, \|\mathbf{w}\|_p^p \leq \gamma$$
$$\iff \min_{\mathbf{w}} L(\mathbf{w}) = \sum_{n=1}^{N} (f(\mathbf{x}_n; \mathbf{w}) - t_n)^2 + \lambda \|\mathbf{w}\|_p^p$$

2. **10 Points.** Discuss the relationship between hyperparameters λ and γ.

Problem 5 (20 Points): Convergence Analysis of Gradient Descent

In the class, I mentioned that using gradient descent to solve linear regression can converge to the optimal parameters under certain conditions. This last problem is about generalizing gradient descent to the *convex* function. The formal conditions and convergence property is in the following theorem. **Prove this theorem!**

Theorem 1 *Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and that its gradient ∇f is Lipschitz continuous with constant $L > 0$. Then if we run gradient descent for K iterations with a fixed learning rate $0 < \alpha \leq \frac{1}{L}$, it will produce a solution $f^{(K)}$ which satisfies:*

$$f(\mathbf{x}^{(K)}) - f(\mathbf{x}^*) \leq \frac{1}{2\alpha K} \|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^2, \tag{6}$$

where $f(\mathbf{x}^\star)$ is the optimal value. Intuitively, this means that gradient descent is guaranteed to converge and that it converges with rate $O(1/K)$.

Preliminaries:

1. A convex function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies that: For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x})$.
2. The gradient ∇f is L-Lipschitz continuous, which implies that:

$$\|f(\mathbf{y}) - f(\mathbf{x})\|_2 \leq L\|\mathbf{y} - \mathbf{x}\|_2 \text{ or } \nabla^2 f(\mathbf{x}) \leq L\mathbf{I}.$$

   Furthermore, by expending f around $f(\mathbf{x})$ we have

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}\nabla^2 f(\mathbf{x})\|\mathbf{y} - \mathbf{x}\|_2^2$$
$$\leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}L\|\mathbf{y} - \mathbf{x}\|_2^2$$

3. The iterative process of gradient descent is:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \nabla f(\mathbf{x}^{(k)}).$$

Hints: (1) **8 Points.** Prove that $f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}) - (1 - \frac{1}{2}L\alpha)\alpha \cdot \|\nabla f(\mathbf{x}^{(k)})\|_2^2$; By applying $\alpha \leq \frac{1}{L}$, it indicates $f(\mathbf{x}^{(k+1)}) \leq f(\mathbf{x}^{(k)}) - \frac{1}{2}\alpha \cdot \|\nabla f(\mathbf{x}^{(k)})\|_2^2$ and f is a strictly decreasing function.

(2) **8 Points.** Prove that: $f(\mathbf{x}^{(k+1)}) - f(\mathbf{x}^\star) \leq \frac{1}{2\alpha}\big(\|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_2^2 - \|\mathbf{x}^{(k)} - \alpha\nabla f(\mathbf{x}^{(k)}) - \mathbf{x}^\star\|_2^2\big)$. By applying $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha\nabla f(\mathbf{x}^{(k)})$, it indicates that $f(\mathbf{x}^{k+1}) - f(\mathbf{x}^\star) \leq \frac{1}{2\alpha}\big(\|\mathbf{x}^{(k)} - \mathbf{x}^\star\|_2^2 - \|\mathbf{x}^{(k+1)} - \mathbf{x}^\star\|_2^2\big)$.

(3) **4 Points.** Prove that $\sum_{k=1}^{K}\big(f(\mathbf{x}^{(k)}) - f(\mathbf{x}^\star)\big) \leq \frac{1}{2\alpha}\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|_2^2$. Using that f decreases on every iteration, we can reach Inequality (6) and thus complete the proof.