Contents

TD 3 Regularization and Optimization

Exercise 3.1 Gradients of Regularization Terms

Question 3.1.a) Linear Regression without Regularization

The objective function for linear regression is to minimize the squared error:

$$\mathcal{L}(w) = \|y - Xw\|^2 = (y - Xw)^\top (y - Xw),$$

where:

- X is the $N \times d$ design matrix (inputs),
- y is the $N \times 1$ vector of outputs,
- w is the $d \times 1$ vector of weights.

Convexity Justification:

1. The squared norm $\| \cdot \|^2$ is a convex function because $\| \cdot \|$ is convex, and squaring a convex, non-negative function preserves convexity.
2. The argument $y - Xw$ is a linear transformation of w, and the composition of a convex function ($\| \cdot \|^2$) with a linear function preserves convexity.
3. The objective $\mathcal{L}(w) = \|y - Xw\|^2$ is therefore convex as the composition of $\| \cdot \|^2$ with a linear function.

Since the objective is convex, any critical point obtained by setting $\nabla_w \mathcal{L}(w) = 0$ is a **global minimum**.

Alternative:

- The Hessian of the objective function is given by:

$$H(w) = \nabla_w^2 \mathcal{L}(w) = 2X^\top X.$$

- The matrix $X^\top X$ is positive semi-definite (as it is a Gram matrix), so $H(w)$ is positive semi-definite.

- Therefore, the loss $\mathcal{L}(w)$ is a convex function, and the solution obtained by setting $\nabla_w \mathcal{L}(w) = 0$ is guaranteed to be a **global minimum**.

Solution:

1. Expand the loss:
$$\mathcal{L}(w) = y^\top y - 2y^\top Xw + w^\top X^\top Xw.$$
2. Compute the gradient with respect to w:
$$\nabla_w \mathcal{L}(w) = -2X^\top y + 2X^\top Xw.$$
3. Set the gradient to zero:
$$-2X^\top y + 2X^\top Xw = 0.$$
4. Solve for w:
$$w = \left(X^\top X\right)^{-1} X^\top y.$$

Case $d = 1$: For $d = 1$, the matrix $X^\top X$ reduces to a scalar:

$$X^\top X = \sum_{i=1}^{N} x_i^2, X^\top y = \sum_{i=1}^{N} x_i y_i.$$

Thus, the solution simplifies to:

$$w = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}.$$

Question 3.1.b) Ridge-regularized Linear Regression

The Ridge regression loss includes a regularization term:

$$\mathcal{L}(w) = \|y - Xw\|^2 + \lambda \ \|w\|_2^2.$$

Here, $\|w\|_2^2 = w^\top w$, and $\lambda > 0$ is the regularization parameter.

Convexity Justification:

1. From part (a), we already established that the term $\|y - Xw\|^2$ is convex.
2. The regularization term $\|w\|_2^2 = w^\top w$ is a quadratic function, which is convex because:
   - The quadratic form $w^\top w$ has a Hessian equal to the identity matrix I, which is positive definite.
   - Therefore, $\|w\|_2^2$ is convex.
3. The overall objective is a sum of two convex functions:
$$\mathcal{L}w = \|y - Xw\|^2 + \lambda \ \|w\|_2^2,$$
   and the sum of convex functions is convex.

Thus, the objective is convex, and any solution obtained by setting $\nabla_w \mathcal{L}w = 0$ is a **global minimum**.

Justification of Minimum:

- The Hessian of the objective function is:

$$H(w) = \nabla_w^2 \mathcal{L}(w) = 2X^\top X + 2\lambda I.$$

- The term $X^\top X$ is positive semi-definite, and ad ding λI (with $\lambda > 0$) ensures that $H(w)$ is strictly positive definite.
- Therefore, $\mathcal{L}(w)$ is strictly convex, and the solution obtained by setting $\nabla_w \mathcal{L}(w) = 0$ is guaranteed to be a **global minimum**.

Solution:

1. Expand the loss:

$$\mathcal{L}(w) = y^\top y - 2y^\top Xw + w^\top X^\top Xw + \lambda w^\top w.$$

2. Compute the gradient with respect to w:

$$\nabla_w \mathcal{L}(w) = -2X^\top y + 2X^\top Xw + 2\lambda w.$$

3. Set the gradient to zero:

$$-2X^\top y + 2X^\top Xw + 2\lambda w = 0.$$

4. Solve for w:

$$w = \left(X^\top X + \lambda I\right)^{-1} X^\top y.$$

Case $d = 1$: For $d = 1$, $X^\top X = \sum_{i=1}^N x_i^2$, and the regularization term ad ds λ. The solution becomes:

$$w = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2 + \lambda N}.$$

Exercise 3.2 Weight Shrinkage in Ridge-regularized Linear Regression

Gradient of Ridge-regularized Linear Regression

From Exercise 3.1, the gradient of the Ridge-regularized loss function is:

$$\mathcal{L}(w) = \|y - Xw\|^2 + \lambda \ \|w\|_2^2.$$

The gradient with respect to w is:

$$\nabla_w \mathcal{L}(w) = -2X^\top y + 2X^\top Xw + 2\lambda w.$$

Simplified:

$$\nabla_w \mathcal{L}(w) = 2X^\top (Xw - y) + 2\lambda w.$$

Gradient Descent Update Step

Gradient descent updates the weights using the rule:

$$w_{k+1} = w_k - \eta \nabla_w \mathcal{L}(w),$$

where:

- w_k is the weight vector at iteration k,

- $\eta > 0$ is the learning rate,
- $\nabla_w \mathcal{L}(w)$ is the gradient of the loss.

Substituting the gradient:

$$w_{k+1} = w_k - \eta[2X^\top (Xw_k - y) + 2\lambda w_k].$$

Simplify:

$$w_{k+1} = w_k - 2\eta X^\top (Xw_k - y) - 2\eta\lambda w_k.$$

Factorize w_k where relevant:

$$w_{k+1} = (1 - 2\eta\lambda)w_k - 2\eta X^\top (Xw_k - y).$$

Geometric Decay (Weight Shrinkage)

The term $(1 - 2\eta\lambda)w_k$ introduces a multiplicative decay factor to w_k. Here's the interpretation:

1. **Shrinkage Factor:**
   - The scalar factor $(1 - 2\eta\lambda)$ reduces the magnitude of w_k at each iteration.
   - If $0 < 2\eta\lambda < 1$, then $1 - 2\eta\lambda < 1$, resulting in a gradual "shrinkage" of the weights.
2. **Geometric Decay:**
   - Ignoring the contribution from the data term $-2\eta X^\top (Xw_k - y)$, the weights evolve as:

     $$w_{k+1} \approx (1 - 2\eta\lambda)w_k.$$

   - Over k iterations, this corresponds to:

     $$w_k \approx (1 - 2\eta\lambda)^k w_0.$$

   - This is an exponential (geometric) decay, where $(1 - 2\eta\lambda)^k$ determines the rate of weight reduction.
3. **Effect of Regularization:**
   - The λw_k term encourages smaller weights, as it penalizes large values of w_k.
   - This prevents overfitting by gradually reducing the influence of less significant features in the model.

Exercise 3.3 Lasso Regularization

Question 3.3.a) Lasso-regularized Linear Regression

1. **Full Formula of the Lasso Loss:** The Lasso loss function is given by:

   $$\mathcal{L}(w) = \|y - Xw\|^2 + \lambda \, \|w\|_1,$$

   where:
   - $\|y - Xw\|^2$ is the least-squares error term.
   - $\|w\|_1 = \sum_{d=1}^{D} |w_d|$ is the L_1 norm of the weight vector w, which is used for regularization.
   - $\lambda > 0$ is a regularization parameter controlling the strength of the penalty on the magnitude of the weights.

2. **Justification of Convexity:** We need to prove that the Lasso loss is convex. The total loss consists of two terms:
   - The first term, $\|y - Xw\|^2$, is convex because it is a quadratic function of w, which is convex in linear models.
   - The second term, $\lambda \|w\|_1 = \lambda \sum_{d=1}^{D} |w_d|$, is convex because the absolute value function $|w_d|$ is convex, and the sum of convex functions is also convex.
3. **Deriving the Gradient:** The gradient of the Lasso loss with respect to the weights w is:

   $$\nabla \mathcal{L}(w) = 2X^\top (Xw - y) + \lambda \ \mathrm{sign}(w),$$

   where:
   - $2X^\top (Xw - y)$ is the gradient of the least-squares error term (it's a linear function of w).
   - $\lambda \ \mathrm{sign}(w)$ is the (sub)gradient of the L_1 regularization term, where $\mathrm{sign}(w_d)$ is the (sub)gradient of $|w_d|$:

     $$\mathrm{sign}(w_d) = 1 \text{ if } w_d > 0 - 1 \text{ if } w_d < 0$$

     This non-differentiability at $w_d = 0$ is crucial to understanding the problem.
4. **Problem with Finding an Exact Solution:**
   - The issue with finding an exact solution arises due to the presence of the $\mathrm{sign}(w)$ term in the gradient.
   - Specifically, the gradient is not well-defined at $w_d = 0$ because the subgradient of $|w_d|$ at zero is not unique — it can take any value between -1 and 1.
   - As a result, the Lasso loss is not differentiable at $w_d = 0$, and we cannot directly solve for the weights in a closed-form expression.
   - This non-differentiability makes it impossible to find an exact analytical solution for the weights. Instead, iterative optimization methods (such as gradient descent or subgradient methods) are typically used to find an approximate solution.

Question 3.3.b) Gradient Descent Update for Lasso

1. **Gradient Update Step:** Using gradient descent, the update rule for the weights w is:

   $$w_{k+1} = w_k - \eta \nabla \mathcal{L}(w_k),$$

   where $\eta > 0$ is the learning rate, and $\nabla \mathcal{L}(w_k)$ is the gradient of the Lasso loss. Substituting the gradient expression derived in part (a), we get:

   $$w_{k+1} = w_k - \eta(2X^\top (Xw_k - y) + \lambda \ \mathrm{sign}(w_k)).$$

   This can be rewritten as:

   $$w_{k+1} = w_k - 2\eta X^\top (Xw_k - y) - \eta\lambda \ \mathrm{sign}(w_k).$$

2. **Explanation of Shrinking:** The update step contains the term $-\eta\lambda \ \mathrm{sign}(w_k)$, which has a shrinking effect on the weights:
   - The $\mathrm{sign}(w_k)$ term reduces the magnitude of each weight w_k by a factor of λ at each iteration, causing the weights to shrink.
   - The shrinkage can drive some of the weights exactly to zero, effectively performing feature selection by excluding irrelevant features from the model.

- This shrinkage behavior is what makes Lasso particularly useful in situations where we expect that many of the features are irrelevant (i.e., the corresponding weights should be zero).

The shrinking effect in Lasso is different from Ridge regression (L_2 regularization), where the weights are only shrunk towards zero but never exactly zero. In contrast, Lasso can force weights to become exactly zero due to the L_1 norm, leading to sparse solutions. This characteristic is useful in feature selection and high-dimensional datasets where many features are irrelevant.

Exercise 3.4 Maximum A Posteriori Estimation (MAP)

Question 3.4.a) MAP Estimation with Exponential Prior

Recall of MAP Estimation (General Framework)

In general, Maximum A Posteriori (MAP estimation is used to find the value of a parameter θ that maximizes the posterior distribution $p(\theta \mid X)$. According to Bayes' theorem, the posterior is proportional to the likelihood of the data given the parameter, $p(X \mid \theta)$, and the prior distribution $p(\theta)$:

$$p(\theta \mid X) \propto p(X \mid \theta) p(\theta)$$

To find the MAP estimate, we maximize the logarithm of the posterior, since the logarithm is a monotonic function:

$$\hat{\theta}_{\text{MAP}} = \operatorname{argmax}_{\theta} \log p(X \mid \theta) + \log p(\theta)$$

2. MAP Estimation for n Data Points (Gaussian Likelihood + Exponential Prior)

For our specific problem, the likelihood of n data points $(x_1, ..., x_n)$ is given by the Gaussian distribution:

$$p(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

The likelihood for all n data points is the product of individual likelihoods:

$$p(X \mid \mu) = \prod_{i=1}^{n} p(x_i \mid \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

The prior for μ is exponential:

$$p(\mu) = \lambda \exp(-\lambda\mu), \mu \geq 0$$

Thus, the posterior distribution is:

$$p(X \mid \mu) p(\mu) = \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right) \exp(-\lambda\mu)$$

Taking the logarithm:

$$\log p(\mu \mid X) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - \lambda\mu = -\frac{1}{2\sigma^2} \|X - \mu\|^2 - \lambda\mu I_d$$

Hence, the MAP estimate for μ is:

$$\mu_{\text{MAP}} = \text{argmin}_\mu \frac{1}{2\sigma^2} \|X - \mu\|^2 + \lambda\mu$$

3. Solving the Argmin (Maximizing the Log Posterior)

We aim to minimize:

$$l : \mu \mapsto \frac{1}{2\sigma^2} \|X - \mu\|^2 + \lambda\mu$$

The function is convexe as the sum of a quadratic and a linear function. The minimum is reached when the derivative is zero.

The gradient of l is:

$$\nabla l(\mu) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) + \lambda$$

Hence by setting the gradient to zero, we find:

$$\text{argmin}_\mu l = \frac{1}{n} \sum_{i=1}^{n} x_i + \lambda \frac{\sigma^2}{n}$$

Conclusion

The MAP estimate for the mean μ of the Gaussian distribution is:

$$\mu_{\text{MAP}} = \frac{1}{n} \sum_{i=1}^{n} x_i - \lambda\sigma^2$$

Question 3.4.b) MAP Estimation with Laplace Prior

MAP Estimation for n Data Points (Gaussian Likelihood + Laplace Prior)

For our specific problem, the likelihood of n data points $(x_1, ..., x_n)$ is given by the Gaussian distribution:

$$p(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

The prior for μ is Laplace:

$$p(\mu) = \frac{1}{2b} \exp\left(-\frac{|\mu|}{b}\right), -\infty < \mu < \infty$$

Thus, the posterior distribution is:

$$p(X \mid \mu)p(\mu) = \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right) \cdot \frac{1}{2b} \exp\left(-\frac{|\mu|}{b}\right)$$

Taking the logarithm:

$$\log p(\mu \mid X) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{|\mu|}{b}$$

3. Solving the Argmax (Maximizing the Log Posterior)

We aim to minimize:

$$l : \mu \mapsto \frac{1}{2\sigma^2} \|X - \mu\|^2 + \frac{|\mu|}{b}$$

This function is convex because it is the sum of a quadratic term and a linear absolute value term, which preserves convexity.

The gradient of the function with respect to μ is:

$$\begin{aligned} \nabla l(\mu) &= -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) + \frac{\nabla\ |.|(\mu)}{b} \\ &= \frac{n}{\sigma^2} \left(\mu - \overline{x} + \nabla\ |.|(\mu) \frac{\sigma^2}{nb} \right) \end{aligned}$$

Where:

$$\nabla\ |.|(\mu) = \begin{cases} 1 & \text{if } \mu > 0 \\ -1 & \text{if } \mu < 0 \\ [-1, 1] & \text{if } \mu = 0 \end{cases}$$

This equation involves a non-differentiable term at $\mu = 0$ due to the absolute value in the prior. Thus, solving this explicitly requires more care (e.g., using sub-gradients). We can look at the solution for $\mu \neq 0$ and $\mu = 0$ separately.

- For $\mu \neq 0$, we find:

$$\mu = -\text{sign}(\mu) \frac{\sigma^2}{nb} + \overline{x}$$

- For $\mu = 0$ we only have sub-gradients, so we have $\partial\ |.| = [-1, 1]$ the solution $\mu = 0$ set the gradient to 0, if:

$$\overline{x} \in \nabla\ |.|(\mu) \frac{\sigma^2}{nb} \Leftrightarrow -\frac{\sigma^2}{nb} \leq \overline{x} \leq \frac{\sigma^2}{nb}$$

Conclusion

The MAP estimate for the mean μ of the Gaussian distribution with a Laplace prior is:

$$\mu_{\text{MAP}} = \begin{cases} 0 & \text{if } -\frac{\sigma^2}{nb} \leq \overline{x} \leq \frac{\sigma^2}{nb} \\ \overline{x} - \text{sign}(\mu) \frac{\sigma^2}{nb} & \text{otherwise} \end{cases}$$

Exercise 3.5 Maximum A Posteriori Estimation (MAP) for Regularization

Question 3.5.a) Recovering Ridge Regression with Gaussian Prior

Recall of MAP Estimation (General Framework)

In general, Maximum A Posteriori (MAP) estimation is used to find the value of a parameter θ that maximizes the posterior distribution $p(\theta \mid X)$. According to Bayes' theorem, the posterior is proportional to the likelihood of the data given the parameter, $p(X \mid \theta)$, and the prior distribution $p(\theta)$:

$$p(\theta \mid X) \propto p(X \mid \theta)p(\theta)$$

To find the MAP estimate, we maximize the logarithm of the posterior, since the logarithm is a monotonic function:

$$\hat{\theta}_{\text{MAP}} = \text{argmax}_{\theta} \log p(X \mid \theta) + \log p(\theta)$$

2. Linear Model with Gaussian Noise

We assume a linear model for the output y given the input x:

$$y = w^T x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

The likelihood of the data, given the model parameters w, is:

$$p(Y \mid X, w, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - w^T x_i\right)^2}{2\sigma^2}\right)$$

The prior on the weights w, assuming a Gaussian prior with mean 0 and variance λ^{-1}, is:

$$p(w) = \left(\frac{\lambda}{2\pi}\right)^{\frac{D}{2}} \exp\left(-\frac{\lambda}{2} \|w\|^2\right)$$

Thus, the posterior is:

$$p(w \mid X, Y) \propto p(Y \mid X, w, \sigma^2)p(w)$$

Taking the logarithm:

$$\log p(w \mid X, Y) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 - \frac{\lambda}{2} \|w\|^2$$

To find the MAP estimate, we maximize the log-posterior. This is equivalent to minimizing the following loss function:

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

This is the loss function for **Ridge Regression**, where the first term is the least squares error and the second term is the L2 regularization (ridge penalty), which is the result of the Gaussian prior on the weights.

Conclusion

By assuming a Gaussian prior on the weights w, we recover **Ridge Regression**, which minimizes the least squares error with L2 regularization on the weights.

—

Question 3.5.b) MAP Estimation with Laplace Prior for Weights

MAP Estimation for Linear Model (Laplace Prior for Weights)

Now, we assume that each weight w_d follows a Laplace distribution:

$$w_d \sim \text{Laplace}(0, b) = \frac{1}{2b} \exp\left(-\frac{|w_d|}{b}\right)$$

The likelihood of the data, given the model and weights, remains the same as in part (a):

$$p(Y \mid X, w, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(y_i - w^T x_i\right)^2}{2\sigma^2}\right)$$

The prior for the weights w, assuming each component of w follows a Laplace distribution, is:

$$p(w) = \prod_{d=1}^{D} \frac{1}{2b} \exp\left(-\frac{|w_d|}{b}\right)$$

Thus, the posterior distribution is:

$$p(w \mid X, Y) \propto p(Y \mid X, w, \sigma^2) p(w)$$

Taking the logarithm:

$$\log p(w \mid X, Y) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 - \sum_{d=1}^{D} \frac{|w_d|}{b}$$

Thus, the log-posterior becomes:

$$\log p(w \mid X, Y) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 - \frac{1}{b} \|w\|_1$$

The loss function to minimize is:

$$L(w) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 + \frac{1}{b} \|w\|_1$$

The second term is the L1 regularization (lasso penalty), which is the result of the Laplace prior on the weights.

Conclusion

By assuming a Laplace prior on the weights, we obtain the **Lasso Regression** loss function, which minimizes the least squares error with L1 regularization (Lasso penalty). The L1 penalty encourages sparsity in the weights, meaning that many of the coefficients may be shrunk to zero.

Exercise 3.6 Standardization and regularization