Linear Regression
Variations and Optimizations

Vrishab Anurag Venkataraghavan

Overview

The goal of this task was to implement different optimizations to linear regression and study their effect on performance on the `diamonds.csv` dataset.

Methods Used:

- **First-order methods:** Adam and ADMM (in addition to SGD)
- **Higher-order methods:** Stochastic Quasi-Newton (SQN) and Sub-sampled Hessian-Free (SSHF)

The comparison (final results cell) clearly shows that the **Sub-sampled Hessian-Free (SSHF)** method was the unanimous winner, as it achieved the lowest Mean Squared Error (MSE) on all the training, validation, and test sets. Among the first-order methods, both Adam and ADMM performed significantly better than baseline SGD, which completely failed to converge (However, when testing with smaller learning rates (on the order of 10^{-5}), SGD did perform much better in some cases even surpassing the other models)

1 Stochastic Gradient Descent (SGD)

1.1 Performance and Observed Outcomes

- **Test MSE:** 2.69×10^{12}
- **Observations:** The basic SGD implementation was **highly unstable and failed to converge**. The test MSE was huge, and the convergence plots show the error increasing over epochs. This divergence was caused by the combination of a very small mini-batch size (1) and a fixed learning rate (0.001), and updates were therefore too unstable. Parameters then ”overshot” the minimum and diverged.

2 Adam (Adaptive Moment Estimation)

2.1 Mathematical Methodology

2.1.1 Rationale

Adam improves upon standard SGD by incorporating ”momentum” (the first moment, m_t) to make convergence faster and an average of past squared gradients (the second moment, v_t) to create adaptive learning rates. This makes Adam more robust and less sensitive to the choice of the global learning rate η.

2.1.2 Update Rules

1. **First moment:** $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$
2. **Second moment:** $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
3. **Bias-corrected estimates:** $\hat{m}_t = \frac{m_t}{1-\beta_1^t} \quad , \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t}$
4. **Parameter update:** $\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$

2.2 Performance and Observed Outcomes

- **Test MSE:** 2.99×10^7
- **Observations:** Adam performed decently, fixing the divergence of SGD and converging quickly to a low error and was significantly more stable than SGD. This shows the advantage of adaptive learning rate

3 Alternating Direction Method of Multipliers (ADMM)

3.1 Mathematical Methodology

3.1.1 Rationale

ADMM takes a complex problem and decomposes it into simpler, more manageable subproblems that are solved sequentially (kind of like a GPU). For this problem, the θ-update becomes a standard Ridge Regression problem, and the z-update is quadratic minimization, both of which have closed-form solutions. u enforces consensus between the individual solutions.

3.1.2 Update Rules

Objective: $\min_\theta \frac{1}{2m}\|X_b\theta - y\|^2 + \frac{\lambda}{2}\|\theta\|^2$. ADMM iteratively updates as:

1. **θ-update:** $\theta_{k+1} = (X_b^T X_b + \frac{\lambda}{\rho}I)^{-1}X_b^T(z_k - u_k)$
2. **z-update:** $z_{k+1} = \frac{y/m+\rho(X_b\theta_{k+1}+u_k)}{1/m+\rho}$
3. **u-update (dual variable):** $u_{k+1} = u_k + (X_b\theta_{k+1} - z_{k+1})$

3.2 Performance and Observed Outcomes

- **Test MSE:** 2.80×10^7
- **Observations:** ADMM performed slightly better than Adam, achieving a lower final test error. It otherwise showed comparable performance, still carrying forward the improvements over SGD

4 Stochastic Quasi-Newton (SQN)

4.1 Mathematical Methodology

4.1.1 Rationale

SQN (using L-BFGS) aims for the faster convergence of second-order methods without the computationally expensive task of forming and inverting the Hessian matrix (H). It approximates the inverse Hessian using only gradient information from recent steps, and a memory-efficient two-loop recursion calculates the search direction without ever having to form the matrix

4.1.2 Update Rule

The general update is $\theta_{t+1} = \theta_t + \eta p_t$, where the search direction p_t is an approximation of the Newton step ($-H_t^{-1}g_t$) (which is found via the L-BFGS recursion using the last M pairs of parameter and gradient differences)

4.2 Performance and Observed Outcomes

- **Test MSE:** 2.99×10^7
- **Observations:** SQN performed well, achieving comparable test MSE to Adam and ADMM. The convergence was smooth and effective, but it did not perform significantly better than the first-order methods, despite. Its sophisticated approximation of curvature provided a somewhat similar benefit to Adam's adaptive first-order approach on this dataset

5 Sub-sampled Hessian-Free (SSHF)

5.1 Mathematical Methodology

5.1.1 Rationale

The SSHF method also uses second-order information but is more direct than SQN. It is Hessian-Free because it solves the Newton system $H_t p_t = -g_t$ for the optimal search direction p_t without having to form the full Hessian. This uses the Conjugate Gradient (CG) algorithm - this only requires Hessian-vector products (HVPs). Sub-sampling the data (hence SS) for the HVP calculation further improves efficiency

5.1.2 Update Rule

$$\theta_{t+1} = \theta_t + \eta p_t$$

where p_t is the solution to $H_t p_t = -g_t$ found via CG.

5.2 Performance and Observed Outcomes

- **Test MSE:** 2.06×10^7
- **Observations:** SSHF was the clear best performer, achieving the lowest MSE across all sets. By using more accurate second-order information to find a better search direction, it converged to a better minimum. This method truly highlighted the advantage of a higher-order approach, where SQN happened to somewhat 'fail'

6 Summary Comparison

Table 1: Overall Algorithm Comparison

Algorithm	**Type**	**Test MSE**	**Brief**
SGD	First-Order	2.69×10^{12}	Very unstable, did not converge
Adam	First-Order (Adaptive)	2.99×10^7	Simple execution and fast convergence, improved on SGD
ADMM	First-Order (Splitting)	2.80×10^7	Slightly outperformed Adam
Stochastic QN	Quasi-Second-Order	2.99×10^7	On par with Adam (despite higher-order)
SSHF	**Second-Order**	2.06×10^7	**Best performer**. Shows clear benefit of higher-order methods

Visualization

- **Individual First and Higher Order Comparisons:** Clearly show that SGD is the outlier, diverging instead of converging; also sub-sampled HF is clearly seen to be better performing than Stochastic QN.
- **Overall View:** The "All Methods Comparison" and "Performance Comparison" plots show SGD as the clear outlier, with a much larger order of magnitude of MSE.
- **Without SGD:** These plots allow us to see more closely the performance of the models at a closer scale without SGD skewing the axes. SSHF is clearly the winner here (in train, val and test as it so happens), with the other three being fairly level.

Conclusion

- **Influence of Second Order:** The top performance of SSHF shows that using second-order information can lead to more accurate models compared to first-order methods (even though this is not observed for SQN).
- **Adaptivity:** For first-order methods, the success of Adam and the stability of ADMM compared to basic SGD completely diverging show that simple gradient descent is usually not robust enough for practical applications without significant tuning.

Final Conclusion

- **For highest accuracy:** SSHF
- **For fast and easy solutions:** Adam and ADMM
- **Caution:** Basic SGD should be used cautiously and likely requires additional techniques and hyperparameter tuning to be effective.