DSC 291: Stochastic Optimization
Problem Set 2

Dmitriy Drusvyatskiy

Due: 03/09

Instructions

- Show all work and clearly justify each step.
- State any assumptions you use.
- You may use results from class unless otherwise specified.
- GenAI Policy: Please try doing the exercises by yourself first. Subsequently, if you use GenAI tools, clearly state this in your submitted work.

1 Theory

Problem 1: Consider convex functions $f_1, \ldots, f_k \colon \mathbb{E} \to \mathbb{R}$ and define

$$g(x) = \max_{i=1,\ldots,k} f_i(x).$$

Suppose that each function f_i is β-smooth. The prox-linear algorithm for this problem is the recursion:

$$x_{k+1} = \arg\min_x \left\{ f_i(x_k) + \langle \nabla f_i(x_k), x - x_k \rangle + \frac{\beta}{2}\|x - x_k\|^2 \right\}.$$

Show that the estimate holds:

$$g(x_k) - \min_x g(x) \leq \frac{\beta \|x_0 - x^*\|^2}{2k}$$

where x^* is any minimizer of g.
[**Hint:** Emulate the analogous result for gradient descent.]

Problem 2: Let $f \colon \mathbb{E} \to \mathbb{R}$ be a β-smooth α-strongly convex function. Suppose that we have access to a stochastic gradient $g(x, z)$ satisfying for some constant γ the estimates:

$$\mathbb{E}_z g(x, z) = \nabla f(x) \qquad \text{and} \qquad \mathbb{E}_z \|g(x, z)\|^2 \leq \gamma (f(x) - f^*) \qquad \forall x \in \mathbb{E}.$$

[Note that we showed already that this is the case when $\mathbb{E}_z g(x^*, z) = 0$ for some minimizer x^* of f—the interpolation setting]. Show that the stochastic gradient method $x_{k+1} = x_k - \eta g(x_k, z_k)$ converges linearly:

$$\mathbb{E}f(x_k) - f^* \leq \left(1 - 2\alpha\eta + \frac{\beta\eta^2\gamma}{2}\right)^k (f(x_0) - f^*),$$

as long as $\eta > 0$ is sufficiently small to ensure $2\alpha\eta - \frac{\beta\eta^2\gamma}{2} < 1$. What happens if you now optimize the rate over η?

2 Computation

2.1 Background and Objective

In this assignment, you will study and compare the empirical behavior of:

- stochastic subgradient method (for a nonsmooth problem),
- stochastic gradient descent (SGD) for smooth problems,
- stochastic variance reduced gradient (SVRG).

The goal is to understand how nonsmoothness, batch size, averaging, step-size schedules, condition number, and variance reduction affect optimization speed, stability, and generalization.

This assignment combines derivations, implementation, and experimental analysis.

2.2 Dataset

The dataset consists of:

- $n = 442$ samples,
- $d = 10$ features $x_i \in \mathbb{R}^{10}$,
- a real-valued response $y_i \in \mathbb{R}$.

Load the data via:

```
from sklearn.datasets import load_diabetes
X, y = load_diabetes(return_X_y=True)
```

Randomly split the data into:

- 75% training set,
- 25% validation set.

Fix a random seed and report it.

2.3 Problem Setup

We consider finite-sum problems

$$\min_{w\in\mathbb{R}^d} F(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w).$$

Model A: ℓ_2-Regularized Logistic Regression. Convert responses to labels $y_i \in \{-1, 1\}$. Define

$$f_i(w) = \log\Big(1 + \exp(-y_i x_i^\top w)\Big) + \frac{\lambda}{2}\|w\|^2.$$

Then

$$F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w).$$

For $\lambda > 0$, the function is smooth and strongly convex.

Model B: Least Absolute Deviations (LAD). Define

$$f_i(w) = |x_i^\top w - y_i|, \qquad F(w) = \frac{1}{n}\sum_{i=1}^{n} |x_i^\top w - y_i|.$$

This objective is convex but nonsmooth.

2.4 Part I: Stochastic Subgradient Method for Model B

(a) Subgradients.

1. Compute the subdifferential $\partial|t|$.
2. Show that a stochastic subgradient of F is

$$g_i(w) = \operatorname{sign}(x_i^\top w - y_i)\, x_i.$$

(b) Stochastic Subgradient Method. Consider

$$w_{k+1} = w_k - \eta_k g_{i_k}(w_k), \qquad i_k \sim \operatorname{Unif}(\{1, \dots, n\}).$$

1. Implement stochastic subgradient method.
2. Experiment with step-size schedules:
   - Constant: $\eta_k = \eta$,
   - $\eta_k = \eta_0/\sqrt{k+1}$,
   - $\eta_k = \eta_0/(k+1)$.
3. Plot training and validation loss versus:
   - iterations,
   - gradient evaluations.

2.5 Part II: SGD for Model A

(a) Mini-batch SGD. Let B_k be a batch of size b. Consider

$$w_{k+1} = w_k - \eta_k \frac{1}{b} \sum_{i \in B_k} \nabla f_i(w_k).$$

1. Implement mini-batch SGD.
2. Compare $b \in \{1, 5, 20, 100, n\}$.
3. Plot loss versus gradient evaluations.

Questions.

- How does gradient noise depend on b?
- Which batch size gives fastest decrease per gradient evaluation?
- What happens when $b = n$?

(b) Uniform Averaging. Define

$$\bar{w}_T = \frac{1}{T} \sum_{k=1}^{T} w_k.$$

1. Implement uniform averaging.
2. Compare w_T and $\bar{w}_T$.
3. Plot both losses versus gradient evaluations.

Questions.

- Does averaging reduce variance?
- Does it improve validation performance?

(c) Step Decay with Expanding Epochs. Let $\eta^{(0)}$ and m_0 be initial step size and epoch length. For epoch $s = 0, 1, 2, \ldots$:

- Run SGD for m_s iterations with step size $\eta^{(s)}$,
- Update

$$\eta^{(s+1)} = \gamma \eta^{(s)}, \qquad m_{s+1} = \gamma^{-1} m_s.$$

1. Implement for $\gamma \in \{1/2, 0.8\}$.
2. Compare against constant-step SGD.
3. Plot loss versus gradient evaluations.

2.6 Part III: SVRG for model A

Implement SVRG with inner loop length m and stepsize η in each epoch, plot loss versus gradient evaluations, and compare against best-performing SGD variants. Experiment with the following parameter settings:

1. Choose $\eta \in \{0.05/L, 0.1/L, 0.2/L\}$.
2. Choose $m \in \{\lceil 0.5\kappa \rceil, \lceil \kappa \rceil, \lceil 2\kappa \rceil\}$.

Questions.

- Do you observe linear convergence?
- How sensitive is performance to m?
- Does $m \gg \kappa$ degrade performance?
- Which method performs best under equal computational budget?