# DSC 291: Stochastic Optimization

## Problem Set 1

Dmitriy Drusvyatskiy

Due: January 30

## 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 have trouble or you want a hint from online resources, you can use GenAI. If you do so, please state that you used GenAI in your submitted work.

## 1 Theory

**Some notation:** For any linear operator $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$ on a Euclidean space $\mathbf{E}$ there exists a unique linear operator $\mathcal{A}^*\colon \mathbf{E} \to \mathbf{E}$, which we call the *adjoint*, satisfying $\langle \mathcal{A}x, y\rangle = \langle x, \mathcal{A}^* y\rangle$ for every $x, y \in \mathbf{E}$. We say that $\mathcal{A}$ is *self-asjoint* if equality $\mathcal{A} = \mathcal{A}^*$ holds. In the usual case $\mathbf{E} = \mathbf{R}^d$, if $\mathcal{A}$ is represented as a matrix $A \in \mathbf{R}^{d\times d}$, then $\mathcal{A}^*$ is represented by the transpose $A^\top$. Therefore self-adjoint linear operators exactly correspond to symmetric matrices.

### Problem 1

Given a symmetric positive definite matrix $A \in \mathbf{R}^{d\times d}$, show that the assignment $\langle v, w\rangle_A := \langle Av, w\rangle$ is an inner product on $\mathbf{R}^d$, with the induced norm $\|v\|_A = \sqrt{\langle Av, v\rangle}$.

### Problem 2

Define the function
$$f(x) = \tfrac{1}{2}\langle \mathcal{A}x, x\rangle + \langle v, x\rangle + c$$
where $\mathcal{A}\colon \mathbf{E} \to \mathbf{E}$ is a sel-djoint linear operator, $v$ lies in $\mathbf{E}$, and $c$ is a real number. Derive the equation:
$$\nabla f(x) = \mathcal{A}x + v \qquad \text{and} \qquad \nabla^2 f(x) = \mathcal{A}.$$

### Problem 3

Consider a function $f\colon U \to \mathbf{R}$ and a linear mapping $\mathcal{A}\colon \mathbf{Y} \to \mathbf{E}$ and define the composition $h(x) = f(\mathcal{A}x)$.

1. Show that if $f$ is differentiable at $\mathcal{A}x$, then
$$\nabla h(x) = \mathcal{A}^* \nabla f(\mathcal{A}x).$$
2. Show that if $f$ is twice differentiable at $\mathcal{A}x$, then
$$\nabla^2 h(x) = \mathcal{A}^* \nabla^2 f(\mathcal{A}x)\mathcal{A}.$$

### Bonus Problem

Define the two sets

$$\begin{aligned}\mathbf{R}^n_{++} &:= \{x \in \mathbf{R}^n : x_i > 0 \text{ for all } i = 1, \ldots, n\},\\ \mathbf{S}^n_{++} &:= \{X \in \mathbf{S}^n : X \succ 0\}.\end{aligned}$$

Consider the two functions $f\colon \mathbf{R}^n_{++} \to \mathbf{R}$ and $F\colon \mathbf{S}^n_{++} \to \mathbf{R}$ given by

$$f(x) = -\sum_{i=1}^n \ln x_i \qquad \text{and} \qquad F(X) = -\ln\det(X),$$

respectively. Note, from basic properties of the determinant, the equality $F(X) = f(\lambda(X))$, where we set $\lambda(X) := (\lambda_1(X), \ldots, \lambda_n(X))$.

1. Find the derivatives $\nabla f(x)$ and $\nabla^2 f(x)$ for $x \in \mathbf{R}^n_{++}$.
2. Using the property $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, prove $\nabla F(X) = -X^{-1}$ and $\nabla^2 F(X)[V] = X^{-1}VX^{-1}$ for any $X \succ 0$.

   [**Hint:** To compute $\nabla F(X)$, justify
   $$F(X + tV) - F(X) + t\langle X^{-1}, V\rangle = -\ln\det(I + tX^{-1/2}VX^{-1/2}) + t \cdot \operatorname{tr}(X^{-1/2}VX^{-1/2}).$$
   By rewriting the expression in terms of eigenvalues of $X^{-1/2}VX^{-1/2}$, deduce that the right-hand-side is $o(t)$. To compute the Hessian, observe
   $$(X + V)^{-1} = X^{-1/2}\left(I + X^{-1/2}VX^{-1/2}\right)^{-1} X^{-1/2},$$
   and then use the expansion
   $$(I + A)^{-1} = I - A + A^2 - A^3 + \ldots = I - A + O(\|A\|^2_{op}),$$
   whenever $\|A\|_{op} < 1$. ]
3. Show
   $$\langle \nabla^2 F(X)[V], V\rangle = \|X^{-\frac{1}{2}}VX^{-\frac{1}{2}}\|^2_F$$
   for any $X \succ 0$ and $V \in \mathbf{S}^n$. Deduce that the operator $\nabla^2 F(X)\colon \mathbf{S}^n \to \mathbf{S}^n$ is positive definite.

# 2 Computation

## 2.1 Background and Objective

In this assignment, you will study and compare the empirical behavior of *gradient descent (GD)* and *stochastic gradient descent (SGD)* on a small-scale regression task. The goal is to understand how step size, stochasticity, iterate averaging, and mini-batch size affect optimization speed, stability, and generalization. You will work with the **diabetes dataset**, a standard benchmark in statistical learning, where the task is to predict disease progression one year after baseline. This assignment combines implementation, experimentation, and theoretical interpretation.

## 2.2 Dataset

The dataset consists of:

- $n = 442$ patient samples,
- $d = 10$ real-valued features per patient, i.e. data vectors $x_i \in \mathbf{R}^{10}$ for $i = 1, \ldots, n$
- a target variable representing disease progression, i.e. the label $y_i \in \mathbf{R}$ for $i = 1, \ldots, n$.

The data are already normalized and centered.

**Data access.** You should load the data using `scikit-learn`. In Python, this can be done as follows:

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

You should randomly split the data into:

- 75% training set (approximately 330 samples),
- 25% validation set (approximately 110 samples).

Fix a random seed to ensure reproducibility and report it in your submitted work.

## 2.3 Problem Setup

We consider regularized linear regression. Given training data $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, define the objective

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} (x_i^\top w - y_i)^2 + \frac{\lambda}{2} \|w\|^2, \tag{1}$$

where $\lambda \geq 0$ is a regularization parameter.

Throughout the assignment, you may fix $\lambda$ to a small positive value (e.g. $\lambda = 10^{-3}$), unless otherwise stated.

## 2.4 Part I: Batch Gradient Descent

### (a) Gradient and Smoothness

1. Derive the gradient $\nabla f(w)$ of the objective in Eq. (1).
2. Show that $f$ is convex with $L$-Lipscitz gradient. Express $L$ in terms of the data matrix $X$ and $\lambda$.

### (b) Constant Step-Size GD

Consider batch gradient descent with constant step size $\eta$:

$$w_{k+1} = w_k - \eta \nabla f(w_k). \tag{2}$$

1. Implement GD and run it for a fixed number of epochs. Experiment with step sizes

$$\eta \in \left\{ \frac{0.1}{L}, \frac{0.5}{L}, \frac{1}{L}, \frac{1.5}{L} \right\}.$$

   Plot (2) training loss vs. iteration and (2) validation loss vs. iteration.

**Questions.**

- How does the convergence rate depend on $\eta$?
- What happens when $\eta > 1/L$?

## 2.5 Part II: Stochastic Gradient Descent

### (a) Constant Step-Size SGD

Let $\mathcal{B}_k$ be a mini-batch of size $b$ sampled uniformly from the training set. Define the SGD update:

$$w_{k+1} = w_k - \eta \frac{1}{b} \sum_{i \in \mathcal{B}_k} \nabla \ell_i(w_k), \tag{3}$$

where

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

1. Implement SGD with constant step sizes (experiment with a few stepsizes here).
2. Compare batch sizes

$$b \in \{1, 5, 20, 100, n\}.$$

3. Plot training and validation loss as a function of the number of gradient evaluations.

**Questions.**

- How does gradient noise depend on batch size?
- Which batch size gives the fastest decrease in validation error per gradient evaluation?

### (b) SGD with Iterate Averaging

Define the averaged iterate (Polyak–Ruppert averaging):

$$\bar{w}_T = \frac{1}{T} \sum_{k=1}^{T} w_k. \tag{4}$$

1. Implement SGD with iterate averaging.
2. Compare the performance of:
   - the final iterate $w_T$,
   - the averaged iterate $\bar{w}_T$.

**Questions.**

- How does averaging affect stability?
- Does averaging reduce variance in the objective value?
- Compare averaged SGD to batch GD under a similar computational budget.

## 2.6 Part III: Effect of Batch Size

### (a) Optimization vs. Statistical Error

For different batch sizes:

1. Plot training and validation loss as a function of epochs.
2. Identify regimes dominated by optimization error versus statistical noise.

### (b) Critical Batch Size

Fix a step size $\eta$ and vary the batch size $b$.

1. Identify a batch size beyond which performance gains saturate.
2. Relate your observations to the variance of stochastic gradients.