EXAM: Optimization for Data Science

Robert M. Gower and Alexandre Gramfort

December 30, 2018

This is an open book exam, meaning you can consult any written or printed material. Electronic devices are prohibited. The exam consists of four questions of increasing difficulty. You must justify all of your answers. Good luck!

Ex. 1 — The SAG Algorithm. Consider the optimization problem

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{j=1}^{n} f_j(x), \tag{1}$$

and the following implementation of the SAG algorithm given in Algorithm 1.

Algorithm 1 SAG: Stochastic Average Gradient descent

Initialize $x^0, g_i = 0 \in \mathbb{R}^d$ for $i = 1, \ldots, n$. Choose $\eta > 0$ the stepsize.
for $k = 1, \ldots, T-1$ **do**
  Sample $i_k \in \{1, \ldots, n\}$
  $g_{i_k} = \nabla f_{i_k}(x^k)$
  $G^k = \frac{1}{n}\sum_{j=1}^{n} g_j$
  $x^{k+1} = x^k - \eta G^k$
Output: x^T

Part I

Assume that calculating $\nabla f_{i_k}(x^k)$ costs $O(d)$ operations and that sampling i_k costs $O(1)$. What is the computational cost of a single iteration of Algorithm 1?

Part II

Re-write this implementation of SAG in such a way that the computational cost of a single iteration is $O(d)$.

Ex. 2 — Let $\beta \in \mathbb{R}$ be a variable and $\lambda > 0$ be a given parameter. Consider the soft-thresholding operator given by

$$S_\lambda(\beta) = \begin{cases} \beta - \lambda & \text{if } \lambda < \beta, \\ 0 & \text{if } -\lambda \leq \beta \leq \lambda, \\ \beta + \lambda & \text{if } \beta < -\lambda. \end{cases} \tag{3}$$

Show that the soft-thresholding operator can also be written in a more compact way as

$$S_\lambda(\beta) = \text{sign}(\beta)(|\beta| - \lambda)_+, \tag{4}$$

where $(\alpha)_+ \stackrel{\text{def}}{=} \max\{0, \alpha\}$ and $\text{sign}(\beta)$ is the sign function given by $\text{sign}(\beta) = \begin{cases} 1 & \text{if } \beta \geq 0, \\ -1 & \text{if } \beta < 0. \end{cases}$

Ex. 3 — The SR1 quasi-Newton update. Let $y, s \in \mathbb{R}^d$ be two given vectors and let $H_t \in \mathbb{R}^{d \times d}$ be a given symmetric matrix. Assume that $s^\top (y - H_t s) \neq 0$ and $(y - H_t s) \neq 0$. Consider the SR1 update given by

$$H_{t+1} = H_t + \frac{(y - H_t s)(y - H_t s)^\top}{(y - H_t s)^\top s}. \tag{5}$$

Part I

Show that H_{t+1} is the only matrix that satisfies the secant equation

$$H_{t+1} s = y, \tag{6}$$

and is the result of a symmetric rank one update applied to H_t.

Part II

Assume that H_t is invertible and let $B_t \stackrel{\text{def}}{=} H_t^{-1}$. Using the Sherman–Woodbury formula given by

$$(A + uw^\top)^{-1} = A^{-1} - \frac{A^{-1} u w^\top A^{-1}}{1 + w^\top A^{-1} u}, \tag{7}$$

where $A \in \mathbb{R}^{d \times d}$ is any invertible matrix and $u, w \in \mathbb{R}^d$, show that the inverse $B_{t+1} \stackrel{\text{def}}{=} H_{t+1}^{-1}$ is given by

$$B_{t+1} = B_t \frac{(B_t y - s)(B_t y - s)^\top}{(B_t y - s)^\top y}. \tag{8}$$

Ex. 4 — The Frank-Wolfe Algorithm. Let $f(x)$ be a differentiable convex and L–smooth function over a compact and closed domain $\mathcal{D} \subset \mathbb{R}^n$, that is $f(x)$ satisfies

$$f(y) \geq f(x) + \langle y - x, \nabla f(x) \rangle, \quad \forall x, y \in \mathcal{D}. \tag{12}$$

and

$$f(y) \leq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{L}{2}\|x - y\|_2^2, \quad \forall x, y \in \mathcal{D}. \tag{13}$$

Let C_f be the curvature constant defined by

$$C_f \stackrel{\text{def}}{=} \sup_{\substack{x, s \in \mathcal{D} \\ \gamma \in [0,\, 1] \\ y = x + \gamma(s - x)}} \frac{2}{\gamma^2} \left(f(y) - f(x) - \langle y - x, \nabla f(x) \rangle \right). \tag{14}$$

Part I

Let $D = \sup_{x,s \in \mathcal{D}} \|x - s\|_2$. Prove that

$$C_f \leq L \times D^2. \tag{15}$$

Part II

Let $x^* \stackrel{\text{def}}{=} \min_{x \in \mathcal{D}} f(x)$. Let $h(x) \stackrel{\text{def}}{=} f(x) - f(x^*)$. Let $x^0 \in \mathbb{R}^n$. Consider an iteration of the Frank-Wolfe Algorithm given by

$$s^k = \arg\min_{s \in \mathcal{D}} \left\langle s, \nabla f(x^k) \right\rangle \tag{16}$$

$$\gamma_k = \frac{2}{k+2} \tag{17}$$

$$x^{k+1} = (1 - \gamma_k) x^k + \gamma_k s^k, \tag{18}$$

for $k = 0, \ldots, K$. Show that the iterates x^k satisfy

$$h(x^{k+1}) \leq (1 - \gamma_k) h(x^k) + \frac{\gamma_k^2}{2} C_f. \tag{19}$$

Part III

Induction Argument. Let $h_k > 0$ be a sequence of positive numbers such that

$$h_{k+1} \leq (1 - \gamma_k) h_k + \frac{\gamma_k^2}{2} C_f, \quad \text{for } k = 0, 1, \ldots, \tag{20}$$

where $\gamma_k = \frac{2}{k+2}$. Prove that

$$h_k \leq \frac{2C_f}{k+2}, \quad \text{for } k = 1, 2, \ldots. \tag{21}$$

Use this to show that the Frank Wolfe Algorithm converges sublinearly.