CSIE5410 Optimization Algorithms: HW 2

Yen-Huan Li
CSIE, National Taiwan University

Recall that the Kelly criterion, aka the growth-optimal portfolio, yields the optimization problem

$$x_\star \in \operatorname*{arg\,min}_{x \in \Delta_d} f(x), \quad f(x) \coloneqq -\sum_{i=1}^n w_i \log \langle a_i, x \rangle ,$$

where Δ_d denotes the probability simplex in $\mathbb{R}^d$, w_i are strictly positive numbers satisfying $\sum_i w_i = 1$, and a_i are entrywise non-negative vectors. Suppose that $a_i \neq 0$ for all $1 \leq i \leq n$.

1. (10 points) Show that the function f is 1-smooth relative to the log-barrier ~~and explicitly specify the value of L~~. Recall that the log-barrier is given by

$$h(x) \coloneqq -\sum_{i=1}^d \log x[i],$$

where $x[i]$ denotes the i-th entry of the vector x.

Denote the Bregman divergence associated with h as D_h, i.e.,

$$D_h(y, x) \coloneqq h(y) - [h(x) + \langle \nabla h(x), y - x \rangle] .$$

Consider solving the optimization problem (1) by the following algorithm:

- Let $x_1 = (1/d, \ldots, 1/d) \in \Delta_d$.
- For every $t \in \mathbb{N}$, compute

$$x_{t+1} \in \operatorname*{arg\,min}_{x \in \Delta_d} \langle \nabla f(x_t), x - x_t \rangle + D_h(x, x_t).$$

2. (10 points) Show that for any $x \in \Delta_d$ and $0 \leq \alpha < 1$,

$$f(x_\alpha) \leq f(x) + \frac{\alpha}{1 - \alpha},$$

where $x_\alpha \coloneqq (1 - \alpha)x + \alpha(1/d, \ldots, 1/d)$.

3. (10 points) Suppose that $t > 3d$. Show that the algorithm described above satisfies

$$f(x_{t+1}) - f(x_\star) = O\left(\frac{d \log(t/d)}{t} \right).$$

4. (10 points) Let us now focus on the implementation of the algorithm. Show that

$$x_{t+1} = e \oslash \left[\nabla f(x_t) + (e \oslash x_t) + \lambda e \right],$$

where e denotes the all-ones vector, $\oslash$ denotes the entrywise division, and λ is given by

$$\lambda \in \operatorname*{arg\,min}_{u \in \mathbb{R}} \varphi(u), \quad \varphi(u) \coloneqq u - \sum_{i=1}^{d} \log(u + \nabla f(x_t)[i] + 1/x_t[i]).$$

5. (10 points) Show that the function φ is self-concordant. Therefore, the iterates $(x_t)_{t \in \mathbb{N}}$ can be computed by the algorithms in Chapter 5.2 of *Lectures on Convex Optimization* by Nesterov.

Consider the problem of learning a linear classifier given data $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^d \times \{\pm 1\}$. The ℓ_1-regularized lasso yields the optimization problem

$$w_\star \in \text{arg min}_{w \in \mathbb{R}^d} f(w) + \lambda g(w),$$

for some regularization parameter $\lambda > 0$, where

$$f(x) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \log \left(1 + \mathrm{e}^{-y_i \langle x_i, w \rangle} \right),$$
$$g(w) \coloneqq \|w\|_1.$$

Notice that in HW0, we have proved that the function f is convex and L-smooth for some $L > 0$. Nevertheless, the ℓ_1-norm is not differentiable, so gradient descent does not directly apply.

6. (10 points) Define

$$g_\mu(w) \coloneqq \max_{v \in \mathcal{B}_\infty} \langle w, v \rangle - \frac{\mu}{2} \|v\|_2^2,$$

where $\mathcal{B}_\infty$ denotes the unit ℓ_∞-norm ball. Show that the function g_μ is differentiable and

$$\nabla g_\mu(w)[i] = \begin{cases} 1 & \text{if } w[i] \geq \mu; \\ \frac{w[i]}{\mu} & \text{if } -\mu \leq w[i] \leq \mu; \\ -1 & \text{if } w[i] < -\mu. \end{cases}$$

Here, $\nabla g_\mu(w)[i]$ and $w[i]$ denote the i-th entry of $\nabla g_\mu(w)$ and w, respectively.

7. (10 points) Show that the function g_μ is ($1/\mu$)-smooth.

8. (10 points) Show that

$$g_\mu(w) \leq g(w) \leq g_\mu(w) + \frac{\mu d}{2}.$$

9. (10~~5~~ points) Let $F = f + \lambda g$. Show that by gradient descent and appropriately choosing μ, for any pre-specified time horizon $T \in \mathbb{N}$, we can construct a sequence of iterates $w_1, \dots, w_{T+1}$ such that

$$F(w_{T+1}) - F(w_\star) \leq \frac{\lambda \sqrt{d}}{2\sqrt{T}} \left(\|w_1 - w_\star\|_2^2 + 1 \right) + \frac{L \|w_1 - w_\star\|_2^2}{2T}.$$

10. (5 points) Show that the optimization error bound can be improved if we replace gradient descent with an accelerated gradient descent. Write down the value of μ you choose and the corresponding optimization error bound.

11. (5 points) Indeed, the problem of minimizing $f + \lambda g$ can be directly addressed using the FISTA algorithm proposed by Beck and Teboulle, without approximating g by g_μ. Compare the approach above with FISTA and discuss which is more effective.