Modern Optimization Methods for Big Data Problems
Exercises for Lecture 9
(With Solutions)

Dominik Csiba, Jakub Konečný and Peter Richtárik

February 27, 2017

Contents

1 Samplings and Probability Matrices

In the exercises below we will be working with the following sampling $\hat{S}$ defined on $[n] = \{1, 2, 3, 4, 5\}$. The non-zero probabilities of sets being sampled are the following:

$$\begin{aligned} &\text{Prob}(\hat{S} = \{1, 3\}) = 0.1 \\ &\text{Prob}(\hat{S} = \{2, 3, 4, 5\}) = 0.2 \\ &\text{Prob}(\hat{S} = \{4\}) = 0.4 \\ &\text{Prob}(\hat{S} = \{2, 5\}) = 0.1 \\ &\text{Prob}(\hat{S} = \{1, 4, 5\}) = 0.2 \end{aligned} \tag{1}$$

1.1 Probability vector

Find the probability vector p for the sampling $\hat{S}$ defined in (1)?

Solution:
For each $i \in \{1, 2, 3, 4, 5\}$ we sum up the corresponding probabilities, e.g.

$$\text{Prob}(1 \in \hat{S}) = \text{Prob}(\hat{S} = \{1, 3\}) + \text{Prob}(\hat{S} = \{1, 4, 5\}) = 0.1 + 0.2 = 0.3.$$

Similarly we get the whole vector $p = [0.3, 0.3, 0.3, 0.8, 0.5]$.

1.2 Expected size of a sampling

Check that $\sum_i p_i = \text{E}[|\hat{S}|]$ holds for (1), as given by theory (Lecture 6, Lemma 2).

Solution:
Indeed $\sum_i p_i = \text{E}[|\hat{S}|]$:

$$\begin{aligned} LHS &= 0.3 + 0.3 + 0.3 + 0.8 + 0.5 = \mathbf{2.2} \\ RHS &= 0.1 \cdot 2 + 0.2 \cdot 4 + 0.4 \cdot 1 + 0.1 \cdot 2 + 0.2 \cdot 3 = \mathbf{2.2} \end{aligned}$$

1.3 Probability matrix

Find the probability matrix $\mathbf{P} = \mathbf{P}(\hat{S})$ associated with the sampling $\hat{S}$ defined in (1)?

Solution:
Recall, that $\mathbf{P}_{ij} = \text{Prob}(i \in \hat{S},\ j \in \hat{S})$. With straightforward computation we can show that

$$\mathbf{P} = \begin{pmatrix} 0.3 & 0 & 0.1 & 0.2 & 0.2 \\ 0 & 0.3 & 0.2 & 0.2 & 0.3 \\ 0.1 & 0.2 & 0.3 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.8 & 0.4 \\ 0.2 & 0.3 & 0.2 & 0.4 & 0.5 \end{pmatrix}$$

1.4 Expectation of the indicator matrix

Let $\mathbf{I}$ be the $n \times n$ identity matrix, and let $\hat{S}$ be an arbitrary sampling. Let $\mathbf{I}_{\hat{S}}$ denote the $n \times n$ matrix which is obtained from $\mathbf{I}$ be zeroing out diagonal entries $i \notin \hat{S}$. Show that

$$\mathrm{E}\left[\mathbf{I}_{\hat{S}}\right] = \mathbf{Diag}(p).$$

Solution:
Expectation of a random matrix is obtained by taking expectations of every entry (make sure you understand why!). The off-diagonal elements are trivially zero. The i-th diagonal element is 0 with probability $1 - p_i$ and 1 with probability p_i. Therefore the expectation is $0 \cdot (1 - p_i) + 1 \cdot p_i = p_i$.

1.5 Principal submatrices

Let $\emptyset \neq S \subset [n]$. Let $\mathbf{M}_{\{S\}}$ be the $|S| \times |S|$ submatrix of M involving just the rows and columns in the set S. Prove that if $\mathbf{M}$ is positive definite, then $\mathbf{M}_{\{S\}}$ is also positive definite.

Solution:
By contradiction, if $\mathbf{M}_{\{S\}}$ is not positive definite, then there exist a non-zero vector $y \in \mathbb{R}^{|S|}$ such that $y^\top \mathbf{M}_{\{S\}} y \leq 0$. Define a vector $x \in \mathbb{R}^n$, such that it has entries from y for indices from the set S and zeros elsewhere. Then trivially $x^\top \mathbf{M} x \leq 0$ for a non-zero vector x, which is a contradiction with the positive definiteness of $\mathbf{M}$.

2 Computation of the ESO parameters

Now assume that $f \in C^1(\mathbf{M})$, where $\mathbf{M} = \mathbf{A}^\top \mathbf{A}$, with data matrix $\mathbf{A} \in \mathbb{R}^{4\times 5}$ defined as:

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & -1 & -1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix}. \tag{2}$$

One can easily compute, that

$$\mathbf{M} = \mathbf{A}^\top \mathbf{A} = \begin{pmatrix} 2 & 1 & 1 & 0 & 1 \\ 1 & 3 & 2 & 0 & 1 \\ 1 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 2 \\ 1 & 1 & 0 & 2 & 3 \end{pmatrix}.$$

Recall that the ESO/stepsize parameters $v = (v_1, \ldots, v_n)$ are required to satisfy

$$\mathbf{P} \bullet \mathbf{M} \preceq \mathbf{Diag}\left(v \bullet p\right),$$

where $\mathbf{P}$ is the probability matrix and p is the probability vector (Lecture 7). In the exercises below you will be asked to compute v for some specific samplings $\hat{S}$.

2.1 Uniform serial sampling

Let $\hat{S}$ be the uniform serial sampling. Find $v = (v_1, \ldots, v_n)$.

Solution:
The probability vector for uniform serial sampling is

$$p = [0.2, 0.2, 0.2, 0.2, 0.2]$$

and the probability matrix is

$$\mathbf{P} = \begin{pmatrix} 0.2 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 \\ 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0 & 0.2 \end{pmatrix}.$$

We need to bound $\mathbf{P} \bullet \mathbf{M}$ by a diagonal matrix, i.e.,

$$\mathbf{P} \bullet \mathbf{M} = \begin{pmatrix} 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0 & 0.6 \end{pmatrix} = \mathbf{Diag}\,(v \bullet p)$$

with

$$v = [2, 3, 2, 3, 3],$$

i.e., the diagonal of $\mathbf{M}$.

2.2 General serial sampling

Let $\hat{S}$ be any proper serial sampling. Find $v = (v_1, \ldots, v_n)$.

Solution:
The probability vector for a serial sampling is in general

$$p = [p_1, p_2, p_3, p_4, p_5],$$

where $\sum_i p_i = 1$ and $p_i > 0$. The probability matrix is

$$\mathbf{P} = \begin{pmatrix} p_1 & 0 & 0 & 0 & 0 \\ 0 & p_2 & 0 & 0 & 0 \\ 0 & 0 & p_3 & 0 & 0 \\ 0 & 0 & 0 & p_4 & 0 \\ 0 & 0 & 0 & 0 & p_5 \end{pmatrix}.$$

We need to bound $\mathbf{P} \bullet \mathbf{M}$ by a diagonal matrix, i.e.,

$$\mathbf{P} \bullet \mathbf{M} = \begin{pmatrix} 2p_1 & 0 & 0 & 0 & 0 \\ 0 & 3p_2 & 0 & 0 & 0 \\ 0 & 0 & 2p_3 & 0 & 0 \\ 0 & 0 & 0 & 3p_4 & 0 \\ 0 & 0 & 0 & 0 & 3p_5 \end{pmatrix} = \mathbf{Diag}\left(v \bullet p\right)$$

with

$$v = [2, 3, 2, 3, 3],$$

i.e., the diagonal of $\mathbf{M}$.

2.3 2-nice sampling

Let $\hat{S}$ be the 2-nice sampling. Find $v = (v_1, \ldots, v_n)$.

Solution:
Theory suggests to use v_i defined by

$$v_i = \sum_{j=1}^{d} \left(1 + \frac{(\tau - 1)(\omega_j - 1)}{n - 1}\right) \mathbf{A}_{ji}^2,$$

where $\omega_j = |\{i \ : \ \mathbf{A}_{ji} \neq 0\}|$. In our case $d = 4,\ n = 5,\ \tau = 2$ and $\omega = [5, 2, 4, 2]$. One can now compute the v_i as follows:

$$v_1 = (1+1)\cdot 1^2 + \left(1+\frac{1}{4}\right)\cdot 1^2 + \left(1+\frac{3}{4}\right)\cdot 0^2 + \left(1+\frac{1}{4}\right)\cdot 0^2 = 3.25$$
$$v_2 = (1+1)\cdot 1^2 + \left(1+\frac{1}{4}\right)\cdot 0^2 + \left(1+\frac{3}{4}\right)\cdot (-1)^2 + \left(1+\frac{1}{4}\right)\cdot 1^2 = 5$$
$$v_3 = (1+1)\cdot 1^2 + \left(1+\frac{1}{4}\right)\cdot 0^2 + \left(1+\frac{3}{4}\right)\cdot (-1)^2 + \left(1+\frac{1}{4}\right)\cdot 0^2 = 3.75$$
$$v_4 = (1+1)\cdot 1^2 + \left(1+\frac{1}{4}\right)\cdot (-1)^2 + \left(1+\frac{3}{4}\right)\cdot 1^2 + \left(1+\frac{1}{4}\right)\cdot 0^2 = 5$$
$$v_5 = (1+1)\cdot 1^2 + \left(1+\frac{1}{4}\right)\cdot 0^2 + \left(1+\frac{3}{4}\right)\cdot 1^2 + \left(1+\frac{1}{4}\right)\cdot 1^2 = 5$$

If we did the calculations correctly, v is ready to go and we are done.

If we are not sure, we can check the correctness of v using the definition, i.e., whether

$$\mathbf{P} \bullet \mathbf{M} \preceq \mathbf{Diag}\left(p \bullet v\right).$$

This should hold for v we computed above. If we want to check it, we need to compute all the parts of the expression. Lets us do that. The probability vector for a 2-nice sampling is

$$p = [0.4, 0.4, 0.4, 0.4, 0.4].$$

This can be either directly computed, or derived from the equation $\sum_i p_i = \mathrm{E}[|\hat{S}|]$. The probability matrix is

$$\mathbf{P} = \begin{pmatrix} 0.4 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.4 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.4 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.4 \end{pmatrix}.$$

The probability $\mathbf{P}_{ij} = \mathrm{Prob}(i \in \hat{S} \;\&\; j \in \hat{S})$ can be computed in different ways, one of them is the following: we have 10 possible sets to sample from, while exactly one of them contains both i and j. The next step is to bound

$$\mathbf{P} \bullet \mathbf{M} = \begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 & 0.1 \\ 0.1 & 1.2 & 0.2 & 0 & 0.1 \\ 0.1 & 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 1.2 & 0.2 \\ 0.1 & 0.1 & 0 & 0.2 & 1.2 \end{pmatrix} \preceq \begin{pmatrix} 0.4v_1 & 0 & 0 & 0 & 0 \\ 0 & 0.4v_2 & 0 & 0 & 0 \\ 0 & 0 & 0.4v_3 & 0 & 0 \\ 0 & 0 & 0 & 0.4v_4 & 0 \\ 0 & 0 & 0 & 0 & 0.4v_5 \end{pmatrix} = \mathbf{Diag}\left(v \bullet p\right).$$

If we followed the theory correctly, the following should hold

$$\begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 & 0.1 \\ 0.1 & 1.2 & 0.2 & 0 & 0.1 \\ 0.1 & 0.2 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 1.2 & 0.2 \\ 0.1 & 0.1 & 0 & 0.2 & 1.2 \end{pmatrix} \preceq \begin{pmatrix} 1.3 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1.5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 2 \end{pmatrix},$$

i.e.

$$\begin{pmatrix} 0.5 & -0.1 & -0.1 & 0 & -0.1 \\ -0.1 & 0.8 & -0.2 & 0 & -0.1 \\ -0.1 & -0.2 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & -0.2 \\ -0.1 & -0.1 & 0 & -0.2 & 0.8 \end{pmatrix}$$

should be positive semi-definite. This can be checked using MATLAB or Julia. There is a useful sufficient condition to check whether a matrix is positive semi-definite, which was not covered in the lecture:

Lemma 1. *Let $\mathbf{M} \in \mathbb{R}^{n \times n}$ be a symmetric matrix with non-negative diagonal entries. If $\mathbf{M}$ is diagonally dominant, i.e.,*

$$|\mathbf{M}_{ii}| \geq \sum_{j \neq i} |\mathbf{M}_{ji}|, \quad \forall i \in [n],$$

then $\mathbf{M}$ is positive semi-definite.

Note that this is only a sufficient condition, i.e., there exist positive semi-definite matrices which are not diagonally dominant.

We can easily check that the requirements of Lemma 1 are met with a large margin. One may ask, whether we can choose v to be smaller. Using Lemma 1 we know that the matrix

$$\begin{pmatrix} 0.3 & -0.1 & -0.1 & 0 & -0.1 \\ -0.1 & 0.4 & -0.2 & 0 & -0.1 \\ -0.1 & -0.2 & 0.3 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & -0.2 \\ -0.1 & -0.1 & 0 & -0.2 & 0.4 \end{pmatrix}$$

is also positive semi-definite. One may compute v from setting the last matrix to be equal to $\mathbf{Diag}\,(p \bullet v) - \mathbf{P} \bullet \mathbf{M}$. With straightforward computation we get the vector $v = [2.75, 4, 2.75, 3.5, 4.0]$. We do not know for sure that this is the best v, but it is sufficient. One should note that to compute v using the diagonal dominance argument one has to first compute $\mathbf{M}$. However, this is infeasible for big problems, as it involves $O(n^2)$ inner products.

2.4 Non-uniform sampling

Let $\hat{S}$ be the sampling defined in (1). Find $v = (v_1, \ldots, v_n)$.

Solution:
We do not have a exact formula for this sampling, but we can use some of the theorems cover in class to get v.

Using Theorem 20 (Lecture 7):
The theorem shows that

$$v_i = \sum_{j=1}^{d} \min\{\omega_j, \tau\} \mathbf{A}_{ji}^2$$

satisfies the ESO. We already computed $\omega = [5, 2, 4, 2]$ in the last exercise and $\tau \geq |\hat{S}|$, so $\tau = 4$. Therefore we can compute v as follows

$$\begin{aligned} v_1 &= 4 \cdot 1^2 + 2 \cdot 1^2 + 4 \cdot 0^2 + 2 \cdot 0^2 = 7 \\ v_2 &= 4 \cdot 1^2 + 2 \cdot 0^2 + 4 \cdot (-1)^2 + 2 \cdot 1^2 = 11 \\ v_3 &= 4 \cdot 1^2 + 2 \cdot 0^2 + 4 \cdot (-1)^2 + 2 \cdot 0^2 = 9 \\ v_4 &= 4 \cdot 1^2 + 2 \cdot (-1)^2 + 4 \cdot 1^2 + 2 \cdot 0^2 = 11 \\ v_5 &= 4 \cdot 1^2 + 2 \cdot 0^2 + 4 \cdot 1^2 + 2 \cdot 1^2 = 11 \end{aligned}$$

Using Theorem 17 (Lecture 7):
A more general but a little more difficult approach is to use the formula

$$v_i = \sum_{j=1}^{d} \lambda(C_j \cap \hat{S}) \mathbf{A}_{ji}^2.$$

The complication arises from the fact that we have to compute or bound

$$\lambda(C_j \cap \hat{S}) = \max_{\theta \in \mathbb{R}^n} \{\theta^\top \mathbf{P}(C_j \cap \hat{S})\theta \ : \ \theta^\top \mathbf{Diag}\left(\mathbf{P}(C_j \cap \hat{S})\right)\theta \leq 1\}$$

by something better than $\min\{\omega_j, \tau\}$. We already computed the probability vector and probability matrix in previous exercises, they are

$$p = [0.3, 0.3, 0.3, 0.8, 0.5]$$

and

$$\mathbf{P} = \begin{pmatrix} 0.3 & 0 & 0.1 & 0.2 & 0.2 \\ 0 & 0.3 & 0.2 & 0.2 & 0.3 \\ 0.1 & 0.2 & 0.3 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.8 & 0.4 \\ 0.2 & 0.3 & 0.2 & 0.4 & 0.5 \end{pmatrix},$$

respectively. We can compute the normalized eigenvalues of the matrices $\mathbf{P}(C_j \cap \hat{S})$ using MATLAB or Julia.

Using Theorem 10 (Lecture 7):
We can compute the v directly from the definition. We need to bound

$$\mathbf{P} \bullet \mathbf{M} = \begin{pmatrix} 0.6 & 0 & 0.1 & 0 & 0.2 \\ 0 & 0.9 & 0.4 & 0 & 0.3 \\ 0.1 & 0.4 & 0.6 & 0 & 0 \\ 0 & 0 & 0 & 2.4 & 0.8 \\ 0.2 & 0.3 & 0 & 0.8 & 1.5 \end{pmatrix} \preceq \begin{pmatrix} 0.3v_1 & 0 & 0 & 0 & 0 \\ 0 & 0.3v_2 & 0 & 0 & 0 \\ 0 & 0 & 0.3v_3 & 0 & 0 \\ 0 & 0 & 0 & 0.8v_4 & 0 \\ 0 & 0 & 0 & 0 & 0.5v_5 \end{pmatrix} = \mathbf{Diag}\left(v \bullet p\right).$$

We can use Lemma 1 to compute v in similar fashion as in last exercise. If we choose v so that

$$\mathbf{Diag}\left(v \bullet p\right) - \mathbf{P} \bullet \mathbf{M} = \begin{pmatrix} 0.3 & 0 & -0.1 & 0 & -0.2 \\ 0 & 0.7 & -0.4 & 0 & -0.3 \\ -0.1 & -0.4 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.8 & -0.8 \\ -0.2 & -0.3 & 0 & -0.8 & 1.3 \end{pmatrix},$$

then we know using the diagonal dominance argument that the matrix on the right hand side matrix is positive semi-definite. Now we can compute the v from the equation. We get $v = [3, 5.33, 3.67, 4, 5.6]$.

3 Strong Convexity and Smoothness

3.1 Strong convexity parameter in ridge regression

Find the strong convexity parameter (λ; see Assumption 1, Lecture 6) for the objective function arising in the ridge regression problem:

$$f(x) := \frac{1}{2}\|\mathbf{A}x - b\|^2 + \frac{\mu}{2}\|x\|^2.$$

Solution:

Let's use the definition. Function f is λ-strongly convex if for all $x, h \in \mathbb{R}^n$ it holds that

$$f(x+h) \geq f(x) + \langle \nabla f(x), h \rangle + \frac{\lambda}{2}\|h\|^2.$$

By easy computation (multivariate calculus; chain rule) we have $\nabla f(x) = \mathbf{A}^\top(\mathbf{A}x - b) + \mu x$. Now, plug this in into the definition and we get

$$\begin{aligned}
LHS &= \frac{1}{2}\|\mathbf{A}(x+h) - b\|^2 + \frac{\mu}{2}\|x+h\|^2 \\
&= \frac{1}{2}\|\mathbf{A}x - b\|^2 + \langle \mathbf{A}x - b, \mathbf{A}h \rangle + \frac{1}{2}\|\mathbf{A}h\|^2 + \frac{\mu}{2}\|x\|^2 + \mu\langle x, h\rangle + \frac{\mu}{2}\|h\|^2 \\
&= f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\|\mathbf{A}h\|^2 + \frac{\mu}{2}\|h\|^2 \\
&\geq f(x) + \langle \nabla f(x), h \rangle + \frac{\lambda_{\min}(\mathbf{A}^\top\mathbf{A})}{2}\|h\|^2 + \frac{\mu}{2}\|h\|^2 \\
&= f(x) + \langle \nabla f(x), h \rangle + \frac{\lambda}{2}\|h\|^2 \\
&= RHS
\end{aligned}$$

for

$$\lambda = \lambda_{\min}(\mathbf{A}^\top\mathbf{A}) + \mu.$$

3.2 (*) 1/4-smoothess of logistic loss

Show that the logistic loss $\phi_i(a) = \log(1 + e^{-y_i a})$ function (see Lecture 8) is 1/4-smooth. Assume that the label y_i belongs to the set $\{-1, 1\}$.

Solution:

First, we compute the derivative:

$$\phi_i'(a) = \frac{-y_i}{1 + e^{y_i a}}.$$

Using the definition we want to find constant L_i such that

$$|\phi_i'(a) - \phi_i'(b)| \leq L_i|a - b|$$

for all $a, b \in \mathbb{R}$. Using our computed derivative we have

$$LHS = |\phi_i'(a) - \phi_i'(b)| = \left|\frac{-y_i}{1 + e^{y_i a}} - \frac{-y_i}{1 + e^{y_i b}}\right| = \left|\frac{1}{1 + e^{y_i a}} - \frac{1}{1 + e^{y_i b}}\right|.$$

Using the *mean value theorem* (recall this from a basic calculus course) we have that for a differentiable function f and two points $a, b \in \mathbb{R}$ there exists $c \in [a, b]$ such that $f(b) - f(a) = f'(c)(b - a)$. If we set

$$f(x) = \frac{1}{1 + e^{y_i x}},$$

then

$$f'(c) = \frac{-y_i e^{y_i c}}{(1+e^{y_i c})^2}.$$

If we analyse this expression, we can notice that $|f'(0)| = 1/4$ and that $|f'(c)|$ is increasing until $c = 0$ and decreasing afterwards. This implies, that we can bound $|f'(c)| \leq \frac{1}{4}$, $\forall c \in \mathbb{R}$. Using these arguments we have

$$\begin{aligned} LHS &= \left| \frac{1}{1+e^{y_i a}} - \frac{1}{1+e^{y_i b}} \right| \\ &= |a-b| \cdot \left| \frac{-y_i e^{y_i c}}{(1+e^{y_i c})^2} \right| \\ &\leq \frac{1}{4}|a-b| \\ &= RHS. \end{aligned}$$

Therefore $L_i = 1/4$.

4 Quartz

4.1 Fenchel conjugate of Tikhonov regularizer

Calculate the Fenchel conjugate of the Tikhonov regularizer, i.e., of the function $g(w) = \frac{\lambda}{2}\|w\|_2^2$ (see Lecture 8). That is, compute g^*.

Solution:
Using the definition we have

$$g^*(a) = \max_{s \in \mathbb{R}^d} \{s^\top a - g(s)\} = \max_{s \in \mathbb{R}^d} \left\{ s^\top a - \frac{\lambda}{2}\|s\|^2 \right\}.$$

The maximum over s can be found by setting the derivative to zero (think why!):

$$0 = \frac{d}{ds}\left(s^\top a - \frac{\lambda}{2}\|s\|^2 \right) = a - \lambda s \qquad \Rightarrow \qquad s = \frac{a}{\lambda}.$$

Therefore,

$$g^*(a) = \max_{s \in \mathbb{R}^d} \left\{ s^\top a - \frac{\lambda}{2}\|s\|^2 \right\} = \frac{a^\top a}{\lambda} - \frac{1}{2\lambda}\|a\|^2 = \frac{1}{2\lambda}\|a\|^2.$$

4.2 Importance sampling for Quartz algorithm

The convergence theorem for the Quartz algorithm (see Lecture 8) states

$$t \geq \max_i \left(\frac{1}{p_i} + \frac{v_i}{p_i \lambda \gamma n} \right) \log \left(\frac{P(w^0) - D(\alpha^0)}{\epsilon} \right) \qquad \Rightarrow \qquad \mathrm{E}\left[P(w^t) - D(\alpha^t) \right] \leq \epsilon$$

Deduce the "importance sampling" probabilities p for serial Quartz. In other words, what is the optimal sampling from the class of all serial samplings?

Solution:

Remark: Someone asked for the derivation of importance sampling for NSync; an answer was posted on Learn. The solution for Quartz is similar.

The quantities $\lambda, \gamma, n, \epsilon$ and $P(w^0) - D(\alpha^0)$ are given and the vector v can be computed for serial sampling as

$$v_i = \|\mathbf{A}_{:i}\|^2.$$

The only variable is p. Therefore, our goal is to solve

$$\arg\min_{p \in \mathbb{R}^n_+} \max_{i \in [n]} \left(\frac{1}{p_i} + \frac{v_i}{p_i \lambda \gamma n} \right) \quad \text{s.t.} \quad \sum_{i=1}^{n} p_i = 1.$$

We claim that the solution will satisfy

$$\frac{1}{p_i} + \frac{v_i}{p_i \lambda \gamma n} = \frac{1}{p_j} + \frac{v_j}{p_j \lambda \gamma n} \quad \forall i, j \in [n].$$

Assume this would not be true, i.e., there exist a set of maximizers $M \subset [n],\ M \neq [n]$, such that

$$\frac{1}{p_i} + \frac{v_i}{p_i \lambda \gamma n} \leq \frac{1}{p_j} + \frac{v_j}{p_j \lambda \gamma n} \quad \forall i \in [n],\ \forall j \in M.$$

There exists at least one index k such that

$$\frac{1}{p_k} + \frac{v_k}{p_k \lambda \gamma n} < \frac{1}{p_j} + \frac{v_j}{p_j \lambda \gamma n} \quad \forall j \in M.$$

One may observe, that if we change $p_i \leftarrow p_i + \delta,\ \forall i \in M$ and $p_k \leftarrow p_k - |M|\delta$ for a small enough δ, we will have a better solution. Therefore, we deduce that the solution will satisfy

$$\frac{1}{p_i} + \frac{v_i}{p_i \lambda \gamma n} = \frac{1}{p_j} + \frac{v_j}{p_j \lambda \gamma n} \quad \forall i, j \in [n].$$

It is straightforward to show that

$$p_i \sim v_i + n\lambda\gamma,$$

i.e,

$$p_i = \frac{v_i + n\lambda\gamma}{\sum_{j=1}^{n}(v_j + n\lambda\gamma)}.$$

4.3 (*) Where does the stepsize come from?

In both Quartz and dfSDCA the stepsize θ can be chosen to be any positive value satisfying the inequality

$$\theta \leq \min_i \frac{p_i n\lambda\gamma}{v_i + n\lambda\gamma} \tag{3}$$

In the theorem covered in class, we simply chose θ to be the largest value satisfying the above inequality. Condition (3) appears in one part of the proof, where θ is required to satisfy the inequality

$$\sum_{i=1}^{n} \left(\frac{\theta}{p_i}(v_i + n\lambda\gamma) - n\lambda\gamma \right) \left\| \alpha_i + \nabla\phi_i(\mathbf{A}_i^\top w) \right\|^2 \leq 0. \tag{4}$$

How does (3) follow from (4)?

Solution:
Bounding every summand independently by 0, i.e.

$$\left(\frac{\theta}{p_i}(v_i + n\lambda\gamma) - n\lambda\gamma \right) \left\| \alpha_i + \nabla\phi_i(\mathbf{A}_i^\top w) \right\|^2 \leq 0, \quad \forall i \in [n],$$

we get the required result.

4.4 (*) Can we find a better stepsize?

Looking at (4), we used a loose bound. Can we do better? Can we get equality in (4)? What are the complications and implications?

Solution:
Setting the whole thing to be equal to 0 is going to create the largest possible θ. Let

$$\kappa_i := \|\alpha_i + \nabla\phi_i(\mathbf{A}_i^\top w)\|.$$

The largest θ is

$$\theta = \frac{n\lambda\gamma \sum_{i=1}^n \kappa_i^2}{\sum_{i=1}^n p_i^{-1}(v_i + n\lambda\gamma)\kappa_i^2} \tag{5}$$

The main complication is this: all the gradients appear in the stepsize, so it is inefficient to compute θ this way. On the other hand, *if* we could keep track of this quantity cheaply, we will have a larger stepsize and therefore faster convergence. Also observe, that the importance sampling is going to change!

Are all the complications worth the speed-up? Yes, they might be. The above idea motivated the following paper: [Csiba, Qu and Richtárik: Stochastic dual coordinate ascent with adaptive probabilities, ICML 2015].