lions@epfl

EE-556: MATHEMATICS OF DATA: FROM THEORY TO COMPUTATION
LABORATORY FOR INFORMATION AND INFERENCE SYSTEMS
FALL 2018

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

INSTRUCTOR
PROF. VOLKAN CEVHER

HEAD TA
YA-PING HSIEH

RECITATION 2

We review probability theory and statistics in this recitation. We will also cover basic Stochastic Gradient Descent, an extremely important algorithm for machine learning.

1 Basic Probability and Statistics

This part of the recitation reviews basic probability theory and statistics.

PROBLEM 1: THE MOST USEFUL BOUND IN PROBABILITY THEORY

Recall the definition of probability measure. In this problem, we will rigorously prove **the** most useful bound in probability theory, the Union Bound, stating that given any n events $E_1, E_2, ..., E_n$, we have

$$P(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i).$$

Prove the following statements.

(a) For any two events A and B such that $A \subseteq B$, prove that $P(A) \leq P(B)$.

(b) Prove the union bound $P(\cup_{i=1}^n E_i) \leq \sum_{i=1}^n P(E_i)$.

(Hint: For any two events E_1 and E_2, prove that $P(E_1 \cup E_2) \leq P(E_1) + P(E_2)$. For this, you can use a useful decomposition rule of sets: $E_1 \cup E_2 = (E_1 \cap E_2^c) \cup E_2$. Applying this bound recursively will finish the proof.

To use the recursive arguments, you need to write the union of n events as a union of 2 events.)

2 Randomness in Statistical Learning Problems, and Stochastic Gradient Descent

PROBLEM 3: RECOGNIZING DIFFERENT RANDOMNESS

There are many different randomness in modern data science or machine learning problems. The purpose of this exercise is to help you get a deeper understanding of them.

This course is all about inferring from data, and the data from real world is often random. Besides this intrinsic randomness, another common source of randomness in modern applications is the *randomized algorithms*. It is extremely important that you have a clear picture of what randomness is truly relevant for statistical inference, and what is only for computational purposes.

Consider the Gaussian linear model from Lecture 2: Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $A \in \mathbb{R}^{n \times p}$. We have observations of the form

$$\mathbf{b} = A\mathbf{x}^\natural + \mathbf{w}, \tag{1}$$

where $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 I)$ is the Gaussian noise vector. We aim to solve the maximum likelihood estimator

$$\hat{\mathbf{x}}_{ML} = \arg\min_{x \in \mathbb{R}^p} \frac{1}{2n} \|\mathbf{b} - A\mathbf{x}\|_2^2, \tag{2}$$

where we have normalized the the loss function by the number of measurements n (the number $\frac{1}{2}$ is just for convenience later).

(a) So far the only random component is the noise $\mathbf{w}$. Let $\mathbb{E}_\mathbf{w}$ denote the expectation with respect to the randomness of $\mathbf{w}$. Compute $\mathbb{E}_\mathbf{w} \|\mathbf{b} - A\mathbf{x}^\natural\|_2^2$.

(b) In practice, the measurement matrix A is often random. Assume that the entries of A are independent random variables with mean 0 and variance 1, and are independent of the noise $\mathbf{w}$. Let $\mathbb{E}_A$ denote the expectation with respect to the randomness of A. Show that

$$\frac{1}{n}\mathbb{E}_A\|A\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2 \tag{3}$$

for all $\mathbf{x} \in \mathbb{R}^p$.
(Hint: Let $\mathbf{a_i}^\top$ be a row of A. What is $\mathbb{E}_A|\langle \mathbf{a_i}, \mathbf{x}\rangle|^2$? Can you compute $\mathbb{E}_A\|A\mathbf{x}\|_2^2$ through $\mathbb{E}_A|\langle \mathbf{a_i}, \mathbf{x}\rangle|^2$?)

(c) **(Self-Study)** Show that the following useful basic inequality holds:

$$\|A(\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural)\|_2^2 \leq 2\langle \mathbf{w}, A(\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural)\rangle. \tag{4}$$

(Hint: The maximum likelihood estimator minimizes the loss function, so you can compare the values of the loss function when substituting in $\hat{\mathbf{x}}_{ML}$ and any other $\mathbf{x}$.)

(d) **(Self-Study)** What we ultimately care about is the estimation error: $\mathbb{E}_{A,\mathbf{w}}\|\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\|_2^2$. In view of (b) and (c), one might be tempted to conclude that

$$\mathbb{E}_{A,\mathbf{w}}\|\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\|_2^2 \leq 2\mathbb{E}_{A,\mathbf{w}}\langle \mathbf{w}, \frac{1}{n}A(\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural)\rangle. \tag{5}$$

Please argue carefully why this argument is NOT true.
(Hint: In part (b) we considered a fixed, deterministic $\mathbf{x}$. Is $\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural$ deterministic? If not, what randomness does it depend on?)

(e) We introduce the important Stochastic Gradient Descent (SGD) in the exercise.

Recall Gradient Descent (GD) for minimizing (2):

- Choose $\mathbf{x}_0$ arbitrarily.
- Do $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ for some predetermined step-sizes $\alpha_k > 0$.

Recall also that for (2), the gradient at a point $\mathbf{x}$ is $\nabla f(\mathbf{x}) = \frac{1}{n}A^\top(A\mathbf{x} - \mathbf{b})$.

Consider a randomized algorithm as follows. At the current iterate $\mathbf{x}_k$, an index $i \in \{1, 2, ..., n\}$ is selected uniformly at random. We then replace the gradient at $\mathbf{x}_k$ by the vector $(\langle \mathbf{a_i}, \mathbf{x}_k\rangle - b_i)\mathbf{a_i}$, where $\mathbf{a_i}^\top$ is the i-th row of A and b_i is the i-th element of $\mathbf{b}$. All other steps are the same as GD. Let $\mathbb{E}_{SGD}$ denote the expectation with respect to the randomness of this algorithm. Show that

$$\mathbb{E}_{SGD}\left(\langle \mathbf{a_i}, \mathbf{x}_k\rangle - b_i\right)\mathbf{a_i} = \frac{1}{n}A^\top(A\mathbf{x}_k - \mathbf{b}). \tag{6}$$

That is, the algorithm is a randomized version of Gradient Descent, hence the name.
(Hint: Recall that $\mathbf{a_i}^\top$ is a row of A, and therefore it is a column of $A^\top$.)

Remark: There are many reasons for using SGD instead of GD; we refer the interested students to [1] for a gentle introduction of SGD. Please do bear in mind that SGD is extremely important in practice. Ever heard of deep learning? AlphaGo? Your interest might be piqued if you know that SGD is the go-to algorithm for these state-of-the-art learning machines.

(f) Under the assumptions in (b), show that

$$\mathbb{E}_{A,\mathbf{w}}\left(\langle \mathbf{a_i}, \mathbf{x}\rangle - b_i\right)\mathbf{a_i} = \mathbf{x} - \mathbf{x}^\natural \tag{7}$$

for any $\mathbf{x}$.

(g) Under the assumptions in (b), show that, for any $\mathbf{x}$,

$$\frac{1}{n}\mathbb{E}_{A,\mathbf{w}}A^\top(A\mathbf{x} - \mathbf{b}) = \mathbf{x} - \mathbf{x}^\natural. \tag{8}$$

(Hint: There are many ways of proving this. Combining (e) and (f) gives a very simple proof.)

PROBLEM 4: A CLOSER LOOK AT STOCHASTIC GRADIENT DESCENT

We consider a general finite-sum minimization:

$$\min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \tag{9}$$

where each $f_i : \mathbb{R}^p \to \mathbb{R}$ can be any function. Such template subsumes, in particular, the important Empirical Risk Minimization problems, and we will think of f_i as representing the i-th sample.

(a) Suppose we pick i uniform at random in $\{1, 2, ..., n\}$, and we perform the update:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f_i(\mathbf{x}_k), \tag{10}$$

where α_k's are step-sizes. Prove that $\mathbb{E}_i \nabla f_i(\mathbf{x}) = \nabla F(\mathbf{x})$; that is, (10) is a stochastic version of the gradient descent for minimizing (9).

(b) In (10), we picked only one sample at each iterate. Alternatively, we can fix a positive integer $b \leq n$, and choose b samples at each iterate. To this end, let $\mathcal{J}$ be the set of all subsets of size b for $\{1, 2, ..., n\}$. Pick $J \in \mathcal{J}$ uniformly at random, and consider the update:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \left(\frac{1}{b} \sum_{i \in J} \nabla f_i(\mathbf{x}_k) \right). \tag{11}$$

The resulting algorithm is called the mini-batch SGD, and b is called the batch size.

Prove that $\mathbb{E}\left(\frac{1}{b} \sum_{i \in J} \nabla f_i(\mathbf{x}) \right) = \nabla F(\mathbf{x})$.

References

[1] Léon Bottou *Stochastic Gradient Descent Tricks*, *Neural Networks: Tricks of the Trade*, *page 421-436*, Springer 2012.