Convex Optimization for Machine Learning and Computer Vision

Lecture: Dr. Virginia Estellers
Exercises: Emanuel Laude
Winter Semester 2017/18

Computer Vision Group
Institut für Informatik
Technische Universität München

Weekly Exercises 6

Room: 02.09.023
Friday, 08.12.2017, 09:15-11:00
Submission deadline theory: Monday, 04.12.2017, 10:15, Room 02.09.023

Theory: Consensus Primal-Dual for Sparse SVMs (16 Points)

Exercise 1 (16 Points)**.** In this exercise you are asked to derive the explicit consensus Primal-Dual (PDHG) for sparse binary SVM training. To this end let $\mathcal{I} = \{1, \dots, N\}$ denote a set of training samples that are represented by their feature vectors $x_i \in \mathbb{R}^d$, for $i \in \mathcal{I}$. Each training sample i is associated with a binary class label $y_i \in \{-1, 1\}$. The aim is to estimate a linear classifier parameterized by $w^* \in \mathbb{R}^d, b^* \in \mathbb{R}$ so that $y_i = \text{sign } x_i^\top w^* + b^*$ for most training samples i. Like in the logistic regression task from the previous exercise sheet we assume that the feature vectors are degraded by components containing just noise. In order to jointly train the classifier and select the features we attempt to optimize the model

$$\min_{w \in \mathbb{R}^d, b \in \mathbb{R}} \sum_{i=1}^{N} \ell(w, b; x_i, y_i) + \lambda \|w\|_1, \tag{1}$$

where $\ell(\cdot, \cdot; x_i, y_i)$ is the hinge loss defined according to

$$\ell(w, b; x_i, y_i) := \max\{0, 1 - (\langle x_i, w \rangle + b) y_i\}. \tag{2}$$

For consensus primal-dual we consider an equivalent "lifted" formulation of the above problem: We introduce for each training sample i a "copy" of the classifier variable $w_i = w$, $b_i = b$ and consider the linearly constrained problem

$$\begin{aligned} \min_{\substack{w \in \mathbb{R}^d, b \in \mathbb{R} \\ \{w_i\}, \{b_i\}}} \quad & \sum_{i=1}^{N} \ell(w_i, b_i; x_i, y_i) + \lambda \|w\|_1 \\ \text{subject to} \quad & w_i = w \\ & b_i = b, \forall i. \end{aligned} \tag{3}$$

You are asked to do the following:

1. Bring the consensus model (3) into the standard form

$$\min_x F(Ax) + G(x). \tag{4}$$

Identify the operator A, the functions F and G. What is the optimization variable x? What is its dimension?

Hint: A is a tall matrix of stacked identities.

2. Derive the equivalent Fenchel-Legendre saddle point formulation. Explicitly derive the convex-conjugate F^* of the function F.

Hint: exploit that F "separates" over the training examples, i.e. F is of the form $F(z) := \sum_i f_i(z_i)$, what is the dimension of the variable z? Note that $\dim(z) \gg \dim(x)$.

3. Explicitly derive closed form prox-operators of the functions F^* and G.

4. Explicitly state the PDHG update scheme for sparse SVM optimization.

5. Argue, why this formulation is well suited to distributed (parallel) optimization in large scale machine learning. What are the shortcomings of this formulation? How can the performance potentially be improved?

Hint: Suggested Reader: Boyd et al., Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, Chapters 7.1, 7.2.

Solution. 1. According to the hint, we define the operator $K \in \mathbb{R}^{N(d+1)\times d+1}$ as

$$K := \begin{pmatrix} I \\ I \\ \vdots \\ I \end{pmatrix}, \tag{5}$$

where $I \in \mathbb{R}^{d+1\times d+1}$ is the identity matrix. Define the optimization variable $x \in \mathbb{R}^{d+1}$ as $x := (w, b)$. Then introducing a variable

$$z := \begin{pmatrix} w_1 \\ b_1 \\ w_2 \\ b_2 \\ \vdots \\ w_N \\ b_N \end{pmatrix} \in \mathbb{R}^{N(d+1)},$$

we can compactly write the constraints $w_i = w$ and $b_i = b$, for all $1 \leq i \leq N$ as

$$Kx = z. \tag{6}$$

We identify the function $F : \mathbb{R}^{N(d+1)} \to \mathbb{R}$ as

$$F(z) = \sum_{i=1}^{N} \ell(w_i, b_i; x_i, y_i), \tag{7}$$

and the function $G : \mathbb{R}^{d+1} \to \mathbb{R}$ as

$$G(x) = \lambda \|w\|_1. \tag{8}$$

Then, the problem (3) can be compactly written as

$$\begin{aligned} \min_{x,z} \quad & F(z) + G(x) \\ \text{subject to} \quad & Kx = z, \end{aligned} \tag{9}$$

which is equivalent to the standard form (4).

2. According to the lecture, the equivalent saddle point formulation is given as

$$\min_{x \in \mathbb{R}^{d+1}} \max_{p \in \mathbb{R}^{N(d+1)}} \langle Kx, p \rangle - F^*(p) + G(x). \tag{10}$$

We proceed computing the convex conjugate F^* of F. Since F is separable, the convex conjugate is given as the sum of the conjugates, as

$$F^*(p) := \sum_{i=1}^{N} F_i^*(p_i), \tag{11}$$

where $F_i(z_i) := \max\{0, 1 - (\langle x_i, w_i \rangle + b_i) y_i\}$ and $z_i := (w_i, b_i)$ is a subvector of z. Analogously, the dual variable (Lagrange multiplier) $p_i \in \mathbb{R}^{d+1}$ denotes a sub-vector of p. Both, p_i resp. z_i contain the entries with indices $(d+1)i + 1 \leq j \leq (d+1)(i+1)$ of the vectors $p \in \mathbb{R}^{N(d+1)}$ resp. $z \in \mathbb{R}^{N(d+1)}$.

In order to compute the convex conjugate F_i^* we define the vector $a_i := (x_i, 1) y_i \in \mathbb{R}^{d+1}$ as the product of feature vector $x_i \in \mathbb{R}^d$ and training label $y_i \in \{-1, 1\}$. Then $F_i(w_i, b_i)$ can be written as a composition of a linear function $a_i^\top$ and a scalar nonlinear function $f : \mathbb{R} \to \mathbb{R}$ as

$$F_i(z_i) = \max\{0, 1 - \langle a_i, z_i \rangle\} = f \circ a_i^\top, \tag{12}$$

for $f(t) = \max\{0, 1 - t\}$.

According to the previous exercise sheet, the conjugate of F_i is given via the so called image function, or preimage of f^* w.r.t. a_i, as

$$F_i^*(p_i) := (a_i f^*)(p_i) := \begin{cases} \inf\limits_{\substack{s \in \mathbb{R} \\ a_i s = p_i}} f^*(s) & \text{if } \exists\, s \in \mathbb{R} : a_i s = p_i \\ \infty & \text{otherwise.} \end{cases} \tag{13}$$

It remains to compute the convex conjugate $f^*(s) := \sup_{t\in\mathbb{R}} ts - f(t)$ of $f(t)$. We introduce a substitution $t' = 1 - t$ and obtain:

$$\begin{aligned} f^*(s) &:= \sup_{t\in\mathbb{R}} \; ts - \max\{0, 1-t\} \\ &= \sup_{t'\in\mathbb{R}} \; (1-t')s - \max\{0, t'\} \\ &= \sup_{t'\in\mathbb{R}} \; -t's - \max\{0, t'\} + s \\ &= s + \begin{cases} 0 & \text{if } s \in [-1, 0] \\ \infty & \text{otherwise,} \end{cases} \\ &=: \delta_{[-1,0]}(s) + s. \end{aligned}$$

Overall, the conjugate $F_i^*(p_i)$ is given as

$$\begin{aligned} F_i^*(p_i) &= \begin{cases} \inf\limits_{\substack{s\in\mathbb{R} \\ a_i s = p_i}} \delta_{[-1,0]}(s) + s & \text{if } \exists\, s \in \mathbb{R} : a_i s = p_i \\ \infty & \text{otherwise,} \end{cases} \\ &= \begin{cases} s & \text{if } \exists\, s \in [-1, 0] : a_i s = p_i \\ \infty & \text{otherwise.} \end{cases} \end{aligned}$$

This completes the task.

3. We proceed computing the proximal mapping of F^*. Again, since F^* is separable we can compute the proximal mapping of each summand F_i^* separately.

$$\begin{aligned} \operatorname{prox}_{\sigma F_i^*}(q_i) &:= \operatorname*{arg\,min}_{p_i\in\mathbb{R}^{d+1}} \frac{1}{2\sigma}\|p_i - q_i\|^2 + F_i^*(p_i) \\ &= \operatorname*{arg\,min}_{\substack{s\in[-1,0], p_i\in\mathbb{R}^{d+1}, \\ \text{s.t. } p_i = a_i s}} \frac{1}{2\sigma}\|p_i - q_i\|^2 + s. \end{aligned}$$

We equivalently solve for s only, since p_i is uniquely determined via $p_i := a_i s$ and

$$s^* := \operatorname*{arg\,min}_{s\in[-1,0]} \frac{1}{2\sigma}\|a_i s - q_i\|^2 + s, \tag{14}$$

which is a simple 1d quadratic program. s^* can be obtained by computing the minimum of the 1d-parabola, and clipping the result to the interval $[-1, 0]$.

Then the solution to the proximal mapping can be recovered as

$$\operatorname{prox}_{\sigma F_i^*}(q_i) = a_i s^*. \tag{15}$$

Another, possibility to compute the proximal mapping of F^* is to use Moreau's indentity, and compute the proximal mapping of F instead.

The proximal mapping of G has been derived on a past exercise sheet.

4. Cf. the lecture notes.

5. Since, the proximal mappings can be computed for each training example independently, it is amenable to highly parallel computing architectures such as GPUs. Moreover, not each of the many variables w_i, b_i needs be updated in each iteration. Moreover, they can be updated asynchronously and randomly. This allows for very efficient implementations.

   Suggested Reader: Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications, Chambollle et al., 2017.