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 4

Room: 02.09.023
Friday, 24.11.2017, 09:15-11:00
Submission deadline: Monday, 20.11.2017, 10:15, Room 02.09.023

Theory: Proximal operators and Projections (8 Points)

Exercise 1 (4 Points)**.** Let $A \in \mathbb{R}^{n\times n}$ be orthonormal, meaning that $A^\top A = AA^\top = I$. Let the convex set C be given as

$$C := \{u \in \mathbb{R}^n : \|Au\|_\infty \leq 1\}.$$

Compute a formula for the projection onto C given as

$$\Pi_C(v) := \operatorname{argmin}_{u\in\mathbb{R}^n} \frac{1}{2}\|u - v\|_2^2, \quad \text{s.t. } u \in C.$$

Hint: Show that the ℓ_2-norm of a vector is invariant under a multiplication with an orthonormal matrix A, meaning that $\|u\|_2 = \|Au\|_2$.

Solution. We begin proving the hint:

$$\|Ax\|_2^2 = \langle Ax, Ax\rangle = \langle A^\top Ax, x\rangle = \langle x, x\rangle = \|x\|_2^2.$$

The idea is to rewrite the projection onto the set C in terms of the projection $\Pi_{\tilde{C}}$ onto the unit ball of the ℓ_∞-norm $\tilde{C} := \{x \in \mathbb{R}^n : \|x\|_\infty \leq 1\}$. With the substitution

$$w := Au \iff u = A^\top w$$

and using the hint we obtain:

$$\begin{aligned}
\Pi_C(v) &= \operatorname{argmin}_{\|Au\|_\infty\leq 1} \frac{1}{2}\|v - u\|^2 \\
&= A^\top \operatorname{argmin}_{\|w\|_\infty\leq 1} \frac{1}{2}\|v - A^\top w\|^2 \\
&= A^\top \operatorname{argmin}_{\|w\|_\infty\leq 1} \frac{1}{2}\|A(v - A^\top w)\|^2 \\
&= A^\top \operatorname{argmin}_{\|w\|_\infty\leq 1} \frac{1}{2}\|Av - AA^\top w\|^2 \\
&= A^\top \operatorname{argmin}_{\|w\|_\infty\leq 1} \frac{1}{2}\|Av - w\|^2 \\
&= A^\top \Pi_{\tilde{C}}(Av).
\end{aligned}$$

Exercise 2 (4 Points)**.** Let $f \in \mathbb{R}^n$. Show that the ℓ_1-norm proximity operator of f defined as the solution u of the convex optimization problem

$$\arg\min_{u \in \mathbb{R}^n} \frac{1}{2\lambda}\|u - f\|^2 + \|u\|_1,$$

is given as

$$u \in \mathbb{R}^n, \quad u_i := \begin{cases} f_i + \lambda & \text{if } f_i < -\lambda \\ 0 & \text{if } f_i \in [-\lambda, \lambda] \\ f_i - \lambda & \text{if } f_i > \lambda. \end{cases}$$

Hint: Note that the above optimization problem is decoupled in the sense that one can look for the individual entries u_i of the optimal u separately.

Solution. We begin reformulating the optimality condition

$$0 \in \partial \left(\frac{1}{2\lambda}(u_i - f_i)^2 + |u_i| \right)$$

of the optimal u_i

$$0 = \frac{1}{\lambda}(u_i - f_i) + p, \quad p \in \partial |u_i| := \begin{cases} -1 & \text{if } u_i < 0 \\ [-1, 1] & \text{if } u_i = 0 \\ 1 & \text{if } u_i > 0 \end{cases}$$

$$f_i \in u_i + \begin{cases} -\lambda & \text{if } u_i < 0 \\ [-\lambda, \lambda] & \text{if } u_i = 0 \\ \lambda & \text{if } u_i > 0. \end{cases}$$

Recall that we are looking for a u_i that satisfies the condition above given a fixed f_i. We distinguish the following cases:

1. Assume $f_i \in [-\lambda, \lambda]$. Choosing $u_i := 0$ satisfies the condition above.
2. Assume $f_i > \lambda$. Choosing $u_i := f_i - \lambda$ again satisfies the condition.
3. Assume $f_i < -\lambda$. Choosing $u_i := f_i + \lambda$ is the right choice.

Multinomial Logistic Regression (16 Points)

Exercise 3 (16 Points)**.** In this exercise you are asked to train a linear model for a multiclass classification task with Logistic regression. The idea is as follows: You are given a set of training samples $\mathcal{I} = \{1, \ldots, N\}$ 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 class label $y_i \in \{1, \ldots, C\}$. The aim is to estimate a linear classifier parameterized by $W^* \in \mathbb{R}^{d \times C}, b^* \in \mathbb{R}^C$ so that $y_i = \operatorname{argmax}_{1 \le j \le C} x_i^\top W_j^* + b_j^*$ for most training samples i. Once you have obtained this "optimal" classifier the hope is, that you are able to classify new unseen and unlabeled samples $x \in \mathbb{R}^d$. In machine learning this is called

generalization. For this task you may query your trained model via the classifier rule

$$y = \text{argmax}_{1 \leq j \leq C}\, x^\top W_j^* + b_j^* \tag{1}$$

and y probably is the true class label of x if your model generalizes well.

In order to estimate the model we solve an optimization problem of the form

$$\min_{W \in \mathbb{R}^{d \times C}, b \in \mathbb{R}^C} \frac{1}{N} \sum_{i=1}^{N} \ell(W, b, x_i, y_i) + \frac{\lambda_1}{2} \|W\|_2^2 + \frac{\lambda_1}{2} \|b\|_2^2, \tag{2}$$

where

$$\ell(W, b, x_i, y_i) = -\log \left(\frac{\exp(\langle W_{y_i}, x_i \rangle + b_{y_i})}{\sum_{j=1}^{C} \exp(\langle W_j, x_i \rangle + b_j)} \right) \tag{3}$$

is called the softmax loss. Note that the above problem is smooth and strongly convex and can be solved with gradient descent. In practice however, it may happen, that some features (i.e. components of the vector x_i) do not contain any information about the true class labels, i.e. components that are just noise. In order to filter out the useless features we add the nonsmooth sparsity inducing ℓ_1-norm term on W. So overall we would like to optimize

$$\min_{W \in \mathbb{R}^{d \times C}, b \in \mathbb{R}^C} \frac{1}{N} \sum_{i=1}^{N} \ell(W, b, x_i, y_i) + \frac{\lambda_1}{2} \|W\|_2^2 + \frac{\lambda_1}{2} \|b\|_2^2 + \lambda_2 \|W\|_1. \tag{4}$$

You are asked to do the following:

- Download the toy data template from the homepage
- Implement a proximal gradient descent algorithm to optimize the above objective (Avoid for-loops)
- Make sure that your objective monotonically decreases. Plot the objective values. Stop your code if the difference of two successive iterates is less than 10^{-12}.
- In order to ensure that your derivative is computed correctly you may first optimize the fully differentiable model (2) with MATLABs *fminunc* with the options *'GradObj'*, *'On'* and *'DerivativeCheck'*, *'On'*.
- Iteratively compute the test error in percent, i.e. how many test samples are not classified correctly via the rule (1).
- Play around with different parameter settings for λ_1, λ_2. What do you observe? Can you identify the useless features? Explain why the model generalizes better to unseen test data if you add a sparsity inducing term.

- You may apply your code to the MNIST dataset `http://yann.lecun.com/exdb/mnist/` and see that your are now able to classify handwritten digits.

Solution. We apply the proximal gradient descent scheme to our objective (4). To this end we need compute the partial derivatives $\frac{\partial F(W,b)}{\partial W_{lk}}$ and $\frac{\partial F(W,b)}{\partial b_k}$ of the differentiable part of the objective

$$F(W,b) = \frac{1}{N}\sum_{i=1}^{N} \ell(W,b,x_i,y_i) + \frac{\lambda_1}{2}\|W\|_2^2 + \frac{\lambda_1}{2}\|b\|_2^2.$$

First we observe, that

$$\frac{\partial F(W,b)}{\partial W_{lk}} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell(W,b,x_i,y_i)}{\partial W_{lk}} + \lambda_1 W_{lk}$$

and

$$\frac{\partial F(W,b)}{\partial b_k} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ell(W,b,x_i,y_i)}{\partial b_k} + \lambda_1 b_k.$$

For some class $1 \leq k \leq C$ define

$$h_k(W,b) = \frac{\exp(\langle W_{y_i}, x_i\rangle + b_{y_i})}{\sum_{j=1}^{C}\exp(\langle W_j, x_i\rangle + b_j)}$$

and

$$\mathbf{1}\{y_i = k\} = \begin{cases} 1 & \text{if } y_i = k \\ 0 & \text{otherwise.} \end{cases}$$

Via the one-dimensional chain rule and the quotient rule the partial derivatives of the individual loss terms are given as:

$$\begin{aligned}
&\frac{\partial \ell(W,b,x_i,y_i)}{\partial W_{lk}} \\
=& -\frac{1}{h_{y_i}(W,b)} \cdot \frac{\mathbf{1}\{y_i = k\} \cdot \exp(\langle W_{y_i}, x_i\rangle + b_{y_i}) \cdot x_{il} \cdot \left(\sum_{j=1}^{C}\exp(\langle W_j, x_i\rangle + b_j)\right)}{\left(\sum_{j=1}^{C}\exp(\langle W_j, x_i\rangle + b_j)\right)^2} \\
=& +\frac{1}{h_{y_i}(W,b)} \cdot \frac{\exp(\langle W_{y_i}, x_i\rangle + b_{y_i}) \cdot \exp(\langle W_k, x_i\rangle + b_k) \cdot x_{il}}{\left(\sum_{j=1}^{C}\exp(\langle W_j, x_i\rangle + b_j)\right)^2} \\
=& -\frac{1}{h_{y_i}(W,b)} \cdot \mathbf{1}\{y_i = k\} \cdot x_{il} \cdot h_{y_i}(W,b) + \frac{1}{h_{y_i}(W,b)} \cdot h_{y_i}(W,b) \cdot h_k(W,b) \cdot x_{il} \\
=& (h_k(W,b) - \mathbf{1}\{y_i = k\}) \cdot x_{il}.
\end{aligned}$$

Similarly we obtain for the derivative wrt. b_k:

$$
\begin{aligned}
&\frac{\partial \ell(W, b, x_i, y_i)}{\partial b_k} \\
=& -\frac{1}{h_{y_i}(W, b)} \cdot \frac{\mathbf{1}\{y_i = k\} \cdot \exp(\langle W_{y_i}, x_i \rangle + b_{y_i}) \cdot \left(\sum_{j=1}^{C} \exp(\langle W_j, x_i \rangle + b_j)\right)}{\left(\sum_{j=1}^{C} \exp(\langle W_j, x_i \rangle + b_j)\right)^2} \\
=& +\frac{1}{h_{y_i}(W, b)} \cdot \frac{\exp(\langle W_{y_i}, x_i \rangle + b_{y_i}) \cdot \exp(\langle W_k, x_i \rangle + b_k)}{\left(\sum_{j=1}^{C} \exp(\langle W_j, x_i \rangle + b_j)\right)^2} \\
=& h_k(W, b) - \mathbf{1}\{y_i = k\}.
\end{aligned}
$$