Technische Universität München

Fakultät für Mathematik

Lehrstuhl für Mathematische Optimierung

Optimization Methods for Machine Learning

Michael Ulbrich

July 2017

Planned Contents

The focus of the course is on optimization methods for supervised learning.

Intended contents:

Introduction and Examples

Aspects of Statistical Learning Theory

Supervised Learning Models:

– Classification, especially Support Vector Machines (SVM)

– Regression, especially Neural Networks (deep learning)

Optimization Methods:

– Methods for training SVMs

– Stochastic Gradient and related methods

– Noise reduction (mini-batch methods, gradient aggregation)

– Second order methods (Newton-type approaches)

– Dealing with nonsmoothness

1 Introduction and examples

Machine learning has become a highly important field of research, especially in the context of big data. Many models in supervised learning such as support vector machines or neural networks require training based on data, which calls for suitable nonlinear optimization techniques. This course gives an introduction to modern optimization methods that are well-suited for machine learning tasks. In particular, they a) take into account the specific problem structure that arises in empirical risk minimization, b) are compatible with the results of statistical learning theory, and c) are designed to handle huge amounts of data efficiently. Numerical aspects and illustrative examples will also be part of the lecture.

The field of machine learning is devoted to the development, investigation, and application of methods that enable machines (or computer programs) to learn from data such that they can perform tasks without explicitly being programmed to do so. When new data become available, they can improve their learning autonomously over time. Statistics and optimization are central disciplines that form the mathematical foundation of machine learning. Machine learning plays an important role in, e.g., pattern recognition (such as image, speech), search engines, recommendation systems, and fraud detection.

Machine learning can be divided into different categorized, in particular: supervised learning (classification, regression), unsupervised learning (clustering, sparse coding, dimensionality reduction), and reinforcement learning (finding optimal policies for time-dependent tasks that maximize reward).

This course will focus on optimization methods for supervised learning.

1.1 Basic setup of supervised statistical learning

The general setup of supervised learning is as follows. Given are:

- An input space $\mathcal{X}$ and an output space $\mathcal{Y}$. We set $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.
- A fixed, but unknown Borel probability measure $\mathbb{P}_{\mathcal{X}}$ on $\mathcal{X}$ and a *generator* producing independent samples x distributed according to $\mathbb{P}_{\mathcal{X}}$.
- A *system* (or *supervisor* who, for a given input x, provides an output $y \in \mathcal{Y}$ distributed according to the fixed, but unknown conditional probability $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$. Often, the output corresponding to x is uniquely given by $y = f(x)$ with a fixed but unknown function $f : \mathcal{X} \to \mathcal{Y}$. This corresponds to $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(B|x) = 1$ if $f(x) \in B$ and $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(B|x) = 0$ if $f(x) \notin B$ for any Borel set $B \subset \mathcal{Y}$.
- A *learning machine* (or model) that implements a class of functions $x \in \mathcal{X} \mapsto h(x; w) \in \mathcal{V}$, parametrized by a parameter vector $w \in \mathcal{W}$. Often, there holds $\mathcal{Y} = \mathcal{V}$ or at least that $\mathcal{Y}$ and $\mathcal{V}$ are embedded in a common space.

For the learning task, a *training set* S of N samples $(x^{(i)}, y^{(i)}) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $1 \leq i \leq N$, is available that obey the described sampling and labeling mechanism.

The samples are independently and identically distributed (i.i.d.) according to the probability distribution $\mathbb{P}_{\mathcal{Z}}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ induced by $\mathbb{P}_{\mathcal{X}}$ and $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$:

$$\mathbb{P}_{\mathcal{Z}}(A \times B) = \int_A \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(B|x)\, d\mathbb{P}_{\mathcal{X}}(x).$$

Note that actually the set of samples S is not a set, but a list (or an N-tuple) of samples. There might exist $i \neq j$ with $(x^{(i)}, y^{(i)}) = (x^{(j)}, y^{(j)})$ and although the tuples are the same, they are both different members of the list of samples S.

In the case where the input-output relation is given by a function $f : \mathcal{X} \to \mathcal{Y}$, i.e., when the output y corresponding to the input x is $y = f(x)$, there holds

$$\mathbb{P}_{\mathcal{Z}}(M) = \mathbb{P}_{\mathcal{X}}(\{x \in \mathcal{X} \,:\, (x, f(x)) \in M\}),$$
$$\int_{\mathcal{Z}} \phi(x,y)\, d\mathbb{P}_{\mathcal{Z}}(x,y) = \int_{\mathcal{X}} \phi(x, f(x))\, d\mathbb{P}_{\mathcal{X}}(x).$$

The *task of learning* is to find a parameter $w^* \in \mathcal{W}$ such that $x \mapsto h(x; w^*)$ yields predictions that are consistent with the supervisor's response in the best possible way. The quality of approximation is measured by a loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{W} \to \mathbb{R}$. For fixed $w \in \mathcal{W}$, $\ell(\cdot, \cdot, w)$ is a random variable on $\mathcal{X} \times \mathcal{Y}$.

This leads to the following risk function:

$$R(w) := \int_{\mathcal{Z}} \ell(x, y, w)\, d\mathbb{P}_{\mathcal{Z}}(x, y). \tag{1}$$

We now can formulate the following

Risk minimization problem:

Find $w^* \in \mathcal{W}$ as a solution of

$$\min_{w \in \mathcal{W}} R(w) = \int_{\mathcal{Z}} \ell(x, y, w)\, d\mathbb{P}_{\mathcal{Z}}(x, y). \tag{2}$$

Note that the cost function is just the expected value of $(x, y) \mapsto \ell(x, y, w)$ w.r.t. $\mathbb{P}_{\mathcal{Z}}$.

In the case of the response $x \mapsto y = f(x)$ the goal often times is to approximate f by $h(\cdot; w^*)$ as good as possible. This requires that f and $h(\cdot; w)$ map to a common space. A popular choice is the *least-squares loss function* $\ell(x, y, w) = \frac{1}{2}(h(x; w) - y)^2$.

Remark.

a) Often, we have the structure $\ell(x, y, w) = \tilde{\ell}(y, h(x; w))$. More generally, in almost all cases, there exist functions q, H, and $\hat{\ell}$ such that $h(x; w) = H(q(x; w))$ and $\ell(x, y, w) = \hat{\ell}(y, q(x; w))$.

b) Some authors do not use a parameter space $\mathcal{W}$, but work directly with a set $\mathcal{H}$ of prediction functions $h : \mathcal{X} \to \mathcal{Y}$ and write the loss function in the form $\hat{\ell}(y, h(x))$. If we choose $\mathcal{W} = \mathcal{H}$ and $h(\cdot; w) := w(\cdot)$ for all $w \in \mathcal{W} = \mathcal{H}$ we can recover this setting via $\ell(x, y, w) = \hat{\ell}(y, w(x))$.
There are, however, situations where the loss function cannot be expressed in

terms of the prediction function. This is, e.g., the case for classification by *support vector machines (SVM)* using the *hinge loss* (see below). In this case, the prediction function is

$$h(x;w) = 2I[v^T x + b \geq 0] - 1 = \begin{cases} 1 & \text{if } v^T x + b \geq 0, \\ -1 & \text{if } v^T x + b < 0, \end{cases}$$

where $x \in \mathbb{R}^n$, $w = (v, b) \in \mathbb{R}^n \times \mathbb{R}$, while the hinge loss for the SVM depends on $v^T x + b$:

$$\ell(x, y, w) = \max(0, 1 - y(v^T x + b)).$$

In the remark and throughout the course, the following notation is used:

$$I[\text{true}] = 1 \quad \text{and} \quad I[\text{false}] = 0.$$

The risk minimization problems formulated so far are not numerically tractable since $\mathbb{P}_{\mathcal{Z}}$ is unknown and the above model of learning only provides finitely many samples $(x^{(i)}, y^{(i)})$.

It is natural to approximate the risk

$$R(w) = \mathbb{E}_{\mathcal{Z}}(\ell(\cdot, \cdot; w)) = \int_{\mathcal{Z}} \ell(x, y, w)\, d\mathbb{P}_{\mathcal{Z}}(x, y)$$

by the *empirical risk*

$$R_N(w) = \frac{1}{N} \sum_{i=1}^{N} \ell(x^{(i)}, y^{(i)}, w), \tag{3}$$

which uses the sample $S = (x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})$.

As a computationally tractable problem (at least conceptually) that approximates (2) we consider the

Empirical risk minimization (ERM) problem:

$$\min_{w \in \mathcal{W}} R_N(w) = \frac{1}{N} \sum_{i=1}^{N} \ell(x^{(i)}, y^{(i)}, w). \tag{4}$$

Sometimes, additional knowledge about $w \in \mathcal{W}$ is expressed by a regularization term $Q(w)$, resulting in the following regularized version of (2):

Regularized risk minimization problem:

$$\min_{w \in \mathcal{W}} \int_{\mathcal{Z}} \ell(x, y, w)\, d\mathbb{P}_{\mathcal{Z}}(x, y) + \gamma Q(w), \tag{5}$$

where $\gamma > 0$ is a weight for the regularization. Examples for Q are $Q(w) = \|w\|_2^2$ or $Q(w) = \|w\|_1$ (the latter promotes sparsity of w^*, i.e., that w^* has many zero components).

The regularized empirical version corresponding to (5) is

$$\min_{w\in\mathcal{W}} \underbrace{\frac{1}{N}\sum_{i=1}^{N}\ell(x^{(i)},y^{(i)},w)}_{=R_N(w)} + \gamma Q(w). \tag{6}$$

In empirical risk minimization we can distinguish the following two cases. The choice of prediction functions and of loss functions usually differs in these two cases:

The case where $\mathcal{Y}$ contains only finitely many elements is called *classification*.

The case where $\mathcal{Y}$ contains a continuum of elements is called *regression*.

There are several questions that arise and that are investigated in the field of *learning theory*:

a) If $w_N^* \in \mathcal{W}$, $N \in \mathbb{N}$, are solutions of (4), and $w^* \in W$ is a solution of (2), does there hold

$$R(w_N^*) \to R(w^*),$$
$$R_N(w_N^*) \to R(w^*)$$

in probability as $N \to \infty$? Here, $R(w_N^*)$ are random variables that depend on the random samples.

b) How fast is the rate of convergence and how can it be controlled?

c) What are suitable otpimization methods for ERM?

Analogous questions can be formulated for solutions to the regularized problems (6) and (5).

We will mainly be concerned with c).

For the investigation of numerical methods, it can be helpful to view R_N as an expectation w.r.t. a suitable probability measure. This is achieved by choosing $\mathcal{Z}$ as the underlying space with

$$\mathbb{P}_N(B) = \frac{|\{i\,:\,(x^{(i)},y^{(i)})\in B\}|}{N}.$$

We further note that the choice of the size of the parameter set $\mathcal{W}$ has to be made with care:

Clearly, the larger the parameter set $\mathcal{W}$, the smaller values of $R_N(w_N^*)$ are achievable for fixed sample size N. However, this can lead to *overfitting*, which means that $R_N(w_N^*)$ is small but the *estimation error* $R(w_N^*) - R(w^*)$ is not.

Overfitting can be avoided by suitably controlling the size of $\mathcal{W}$. However, if $\mathcal{W}$ is chosen too restrictive, then it might not be possible to achieve that $R(w^*)$ is small and

in this case $\mathcal{W}$ does not allow to obtain a small *approximation error* $R(w^*) - R_{\min}$. Here $R_{\min}$ is the smallest risk obtainable when not only prediction functions $h(\cdot; w)$, $w \in W$, and the corresponding risk $R(w)$ are considered, but when all possible reasonable prediction functions and the corresponding risks are admitted for the minimization. The situation where the approximation error $R(w^*) - R_{\min}$ is not small can be termed *underfitting*.

1.2 Examples

Examples for classification problems:

- Given a matrix of grey (or black and white) values of a pixel image showing a letter A–Z (or digit 0–9), decide which letter or digit is displayed. The training data would be N instances of such images and labels encoding the displayed letter or number.
- Given an email, decide if it is spam or not. Usually, the input $x \in \mathcal{X}$ corresponding to the email would be obtained by preprocessing, such as using a predetermined dictionary of words or word stems and marking the occurence or counting the relative number of occurences in the mail for all dictionary entries. The training data would be the vectors obtained from N emails, each correctly classified as spam or not spam.
- Other classification tasks include disease prediction (input: medical data collected by examining the patient; output: patient has the specific disease or not) or fraud detection (input: data of bank or credit card transfer activities; output: activities are suspicious or normal behavior).

As we will see, classification problems can also be addressed via regression with the goal to find an approximation for the conditional probability $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$ instead of a "hard" assignment to exactly one class.

Examples for regression problems:

- Given (possibly noisy) measurements $y^{(i)}$ of the output of a device corresponding to inputs $x^{(i)}$, find a function mapping $\mathcal{X}$ to $\mathcal{Y}$ that predicts the input-output behavior of the device in a best possible way. The class of functions over which we optimize is, in our notation, $h(\cdot; w)$, $w \in \mathcal{W}$, which has to be chosen such that it suitably reflects available a priori knowledge about the device. The most well-known regression model is *linear regression*, corresponding to $w = (A, b)$, $h(x; w) = Ax + b$ and the loss function $\ell(x, y, w) = \|h(x; w) - y\|_2^2/2$. Here we assumed $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathbb{R}^m$. Then $w = (A, b)$ with $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$.
- Classification problems can also be tackled by estimating the conditional probability $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}$. For binary classification with $\mathcal{Y} = \{0, 1\}$, $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(\{0\}|x)$ is the probability that the output $y = 0$ occurs when the input is x and similarly for

$\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(\{1\}|x)$. In the case of the input-output relation $y = f(x)$ there then holds $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(\{0\}|x) = 1$ and $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(\{1\}|x) = 0$ if $f(x) = 0$ as well as $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(\{0\}|x) = 0$ and $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(\{1\}|x) = 1$ if $f(x) = 1$.

The approach models $x \mapsto \mathbb{P}_{\mathcal{Y}|\mathcal{X}}(\{1\}|x)$ approximately by a function $h(\cdot; w) : \mathcal{X} \mapsto \mathcal{V} = (0, 1)$ and $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(\{0\}|x)$ by $1 - h(\cdot; w)$. The choice of $h(\cdot; w)$ is typically such that the transition zone of those $x \in \mathcal{X}$ where $h(x; w)$ is neither close to 0 nor 1 is "thin" (with a comparably steep slope of $h(\cdot; w)$ at such x). The probably most prominent such approach is logistic regression, where a linear model $v^T x + b$ sits inside of a logistic (or sigmoid) function $1/(1 + e^{-z})$:

$$h(x; w) = \frac{1}{1 + e^{-(v^T x + b)}}, \quad x \in \mathcal{X} = \mathbb{R}^n, \;\; w = (v, b), \;\; v \in \mathbb{R}^n, \;\; b \in \mathbb{R}.$$

Applying the maximum likelihood paradigm shows that a suitable loss function for the logistic prediction function is given by the log-loss:

$$\ell(x, y, w) = -y \ln(h(x; w)) - (1 - y) \ln(1 - h(x; w)).$$

Examples for loss functions:

- In regression, a standard choice is the least squares loss function

$$\ell(x, y, w) = \tfrac{1}{2} \|h(x; w) - y\|^2.$$

- The probably most natural loss function for classification is the 0-1 loss given by

$$\ell(x, y, w) = I[h(x; w) \neq y],$$

where $\mathcal{V} = \mathcal{Y}$ and $h(\cdot; w) : \mathcal{X} \to \mathcal{Y}$ is the prediction function.

- The *hinge loss* is defined for the case $\mathcal{Y} = \{-1, 1\}$ and has the form

$$\ell(x, y, w) = \max\{0, 1 - yq(x; w)\},$$

where $q(\cdot; w)$ is real valued and the corresponding prediction function is $h(x; w) = 2I[q(x; w) \geq 0] - 1 \in \mathcal{V} = \mathcal{Y}$. It is important in the context of SVM, where usually $q(x; w) = v^T x + b$.

- The *log-loss* for the case $\mathcal{Y} = \{0, 1\}$ is defined by

$$\ell(x, y, w) = -y \ln(h(x; w)) - (1 - y) \ln(1 - h(x; w)),$$

where $h(\cdot; w)$ is a prediction function with values in $\mathcal{V} = (0, 1)$. It is, e.g., used in logistic regression.

- A quite universal method for finding loss functions is the *maximum likelihood* approach. There, the probability density function corresponding to $\mathbb{P}_{\mathcal{Y}|\mathcal{X}}(\cdot|x)$ is approximated by a parametric density function $p(y|x; w)$. We write $y|x$ to highlight that this is a conditional probability density. Note that w is a determinstic parameter. Due to independence of the samples, assuming this distribution, the

joint probability density of $S_y = (y^{(1)}, \dots, y^{(N)})$ given $S_x = (x^{(1)}, \dots, x^{(N)})$ is the *likelihood function*

$$L(w) := p(S_y|S_x; w) = \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}; w).$$

The maximum likelihood approach now chooses w_N^* such that $p(S_y|S_x; w)$ is maximized. Due to the product structure, this is usually done by minimizing

$$-\ln(L(w)) = -\ln(p(S_y|S_x; w) = -\sum_{i=1}^{N} \ln(p(y^{(i)}|x^{(i)}; w)).$$

Comparing with the empirical risk, it is thus natural to choose

$$\ell(x, y, w) = -\ln(p(y|x; w)),$$

which results in

$$R_N(w) = -\frac{1}{N}\ln(L(w)) = -\frac{1}{N}\sum_{i=1}^{N} \ln(p(y^{(i)}|x^{(i)}; w))$$

and

$$R(w) = -\int_{\mathcal{Z}} \ln(p(y|x; w))\, d\mathbb{P}_{\mathcal{Z}}(x, y).$$

If, for instance, we assume $y \sim \mathcal{N}(h(x; w), \sigma^2)$, then we get

$$\begin{aligned}\ell(x, y, w) = -\ln(p(y|x; w)) &= -\ln\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - h(x; w))^2}{2\sigma^2}\right)\right) \\ &= \ln(\sqrt{2\pi}\sigma) + \frac{1}{\sigma^2} \cdot \frac{1}{2}(h(x; w) - y)^2.\end{aligned}$$

Except for the constant offset, which can be dropped in the minimization, and the scaling by $\frac{1}{\sigma^2}$, we obtain the least squares loss function.

Examples for classes of prediction functions

- A general idea that underlies many prediction functions is to use linear functions $x \mapsto v^T x + b$ with parameters $v \in \mathbb{R}^n$ and $b \in \mathbb{R}$ as building blocks. This is done, e.g., in linear regression, where in the scalar-valued case $h(x; w) = v^T x + b$ with $w = (v, b)$ (with obvious extension to the $\mathbb{R}^m$-valued case)

  Support vector machines use $h(x; w) = 2I[v^T x + b \geq 0] - 1$, i.e., $h(x; w) = 1$ if $v^T x + b \geq 0$ and $h(x; w) = -1$, otherwise.

  Logistic regression uses $h(x; w) = \sigma(v^T x + b)$ with $\sigma(z) = 1/(1 + e^{-z})$.

- An often employed trick is to first map x to a feature Hilbert space $\tilde{\mathcal{X}}$ (often with a higher dimension than $\mathcal{X}$) via a mapping $\Phi : \mathcal{X} \to \tilde{\mathcal{X}}$ and then to formulate prediction functions based on $\Phi(x)$ rather than x. This is, e.g., done in classification by SVM if the two classes in $\mathcal{X}$ are far from being approximately separable by a hyperplane. It then turns out that not Φ is required, but only the corresponding kernel $k(x, x') = (\Phi(x), \Phi(x'))_{\tilde{\mathcal{X}}}$, where $(\cdot, \cdot)_{\tilde{\mathcal{X}}}$ is the inner product of $\tilde{\mathcal{X}}$.

- Deep learning uses (deep) neural networks, where deep means that it has multiple layers. Feedforward neural networks consist of several layers, each containing a number of neurons. We introduce a 0th layer, the input layer, with $m_0 = n$ outputs $o_{0,i}(x) = x_i$, $1 \leq i \leq n$, where $x \in \mathbb{R}^n$ is the input to the network. The input $a_l(x) \in \mathbb{R}^{m_l}$ of layer $l \geq 1$ is given by

$$a_l(x) = W_l o_{l-1}(x) + b_l,$$

where the matrix $W_l \in \mathbb{R}^{m_l \times m_{l-1}}$ contains weights and the vector $b_l \in \mathbb{R}^{m_l}$ contains biases. The output $o_l(x) \in \mathbb{R}^{m_l}$ of layer l is obtained as

$$o_{l,j}(x) = \sigma_l(a_{l,j}(x)),$$

where $\sigma_l : \mathbb{R} \to \mathbb{R}$ is an *activation function*. Examples are

$$\text{sign function } \sigma(a) = \operatorname{sgn}(a), \quad \text{threshold function } \sigma(a) = I[a \geq 0],$$
$$\text{sigmoid function } \sigma(a) = \frac{1}{1+e^{-a}}.$$

If the layers are numbered by $0, \ldots, K$, then the output of the network is given by

$$h(x; w) = o_K(x).$$

Here $w = (W_1, b_1, \ldots, W_K, b_K)$ is the parameter, which can be arranged as a vector of length $\sum_{l=1}^{K} m_l(m_{l-1}+1)$.

Sometimes the output layer $l = K$ uses more general activation functions that depend on the whole vector a_K:

$$o_{K,j}(x) = \sigma_{K,j}(a_K(x)).$$

Omitting the layer index K, an example is the *softmax* function

$$\sigma_j(a) = \frac{e^{a_j}}{\sum_{i=1}^{m} e^{a_i}},$$

where $m = m_K$ is the number of neurons in the output layer.

2 Some aspects of statistical learning theory

As described in the introduction, statistical learning uses the concept of empirical risk minimization (ERM)

$$\min_{w \in \mathcal{W}} R_N(w) \left(+ \gamma Q(w) \right)$$

based on a training set S consisting of a list of N i.i.d. samples $(x^{(i)}, y^{(i)})$ to obtain an empirically optimal parameter $w^* \in \mathcal{W}$. Here,

$$R_N(w) = \frac{1}{N} \sum_{i=1}^{N} \ell(x^{(i)}, y^{(i)}, w)$$

with the loss function ℓ.

The true goal, however, is to minimize the expected risk, possibly with a regularization term

$$\min_{w\in\mathcal{W}} R(w) \Big(+\gamma Q(w)\Big),$$

where

$$R(w) = \int_{\mathcal{Z}} \ell(x,y,w)\, d\mathbb{P}(x,y).$$

We first present a theorem that shows the limitations of machine learning. They are inevitable if the class of prediction functions is too large.

For this, we use the following setting (variants are possible):

- $\mathcal{H}$ is the set of all possible prediction functions $h : \mathcal{X} \to \mathcal{Y}$ and we parametrize the prediction functions by themselves (h corresponds to the parameter $h \in \mathcal{W} := \mathcal{H}$).
- We consider binary classification, i.e., $|\mathcal{Y}| = 2$, and use the 0-1-loss $\ell(y, h(x)) = I[h(x) \neq y]$.
- The prediction function is obtained by a learning algorithm A that maps any sample S to a prediction function $A(S) \in \mathcal{H}$.

Then the following theorem holds:

Theorem 2.1 (No Free Lunch Theorem)**.** *Consider a classification problem of the form just described. Let A be any learning algorithm that based on a given sample S generates a prediction function $A(S) \in \mathcal{H}$ (e.g., $A(S)$ could be the solution h_n^* of ERM or of regularized ERM). Let the training set size $N \in \mathbb{N}$ be arbitrarily fixed such that $|\mathcal{X}| > 2N$. Then, there exists a probability measure $\mathbb{P}_{\mathcal{Z}}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ such that:*

a) *There exists a prediction function $h^* : \mathcal{X} \to \mathcal{Y}$ with $R(h^*) = 0$.*

b) *With probability of at least $1/7$ over all samples S of size N (distributed according to $\mathbb{P}_{\mathcal{Z}}^N$), there holds $R(A(S)) \geq 1/8$.*

This essentially shows that for any classification algorithm there exists a distribution for which, although there exists a perfect prediction function $h^* \in \mathcal{H}$, the algorithm performs badly (in the sense of $R(A(S))$ not being small) with a non-neglible probability over the samples as soon as $|\mathcal{X}| > 2N$. This is due to the fact that all possible prediction functions are permitted in the theorem.

A way to escape from this dilemma is to either use a priori knowledge on the distributions $\mathbb{P}_{\mathcal{Z}}$ or to choose a smaller class of prediction functions. It should be not too large, but still contain a prediction function h for which $R(h)$ is small.

For deriving quantitative bounds, we now investigate the *estimation error* $R(w_N^*) - R(w^*)$ in more detail, where w_N^* and w^*, respectively, are global minimizers of R_N

and R, respectively, on $\mathcal{W}$. We focus on the case without regularization and, for greater generality, instead of exact solutions consider *ε-solutions* $w_N^\varepsilon \in \mathcal{W}$ of (4), i.e.,

$$R_N(w_N^\varepsilon) \leq R_N(w_N^*) + \varepsilon.$$

We can estimate the regret as follows:

$$\begin{aligned} R(w_N^\varepsilon) - R(w^*) &= (R(w_N^\varepsilon) - R_N(w_N^\varepsilon)) + \underbrace{(R_N(w_N^\varepsilon) - R_N(w_N^*))}_{\leq \varepsilon} \\ &\quad + \underbrace{(R_N(w_N^*) - R_N(w^*))}_{\leq 0} + (R_N(w^*) - R(w^*)) \\ &\leq (R(w_N^\varepsilon) - R_N(w_N^\varepsilon)) + (R_N(w^*) - R(w^*)) + \varepsilon \\ &\leq \sup_{w \in \mathcal{W}} (R(w) - R_N(w)) + (R_N(w^*) - R(w^*)) + \varepsilon. \end{aligned} \tag{7}$$

In the last step, we made the transition to the sup since w_N^ε depends on the sample and thus can vary depending on the concrete sample. In contrast, w^* does not depend on the sample. However, also w^* depends on the probability distribution $\mathbb{P}_\mathcal{Z}$, which usually is unknown.

For estimating $|R_N(w^*) - R(w^*)|$ we can use the following inequality:

Lemma 2.2 (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_N$ be N independent random variables with $\mathbb{E}(Z_i) = \mu_i$ and $\mathbb{P}\{Z_i \in [a_i, b_i]\} = 1$, $i = 1, \ldots, N$. Then with $\mu = \frac{1}{N}\sum_{i=1}^N \mu_i$ there holds for all $t > 0$:*

$$\mathbb{P}\{|\tfrac{1}{N}\sum_{i=1}^N Z_i - \mu| > t\} \leq 2\exp\left(\frac{-2N^2t^2}{\sum_{i=1}^N (b_i - a_i)^2}\right).$$

In the case $a_i = a$ and $b_i = b$ for all i we can write the bound on the right hand side as $2\exp\left(\frac{-2Nt^2}{(b-a)^2}\right)$.

Proof. This follows immediately from Theorem 2 in [Hoe63] applied once to Z_i and once to $-Z_i$. □

From this lemma, we obtain the following:

Theorem 2.3. *Let $w \in \mathcal{W}$ be fixed and assume that ℓ satisfies $\mathbb{P}_\mathcal{Z}(\{a \leq \ell(x, y, w) \leq b\}) = 1$ with constants $a < b$. Then with probability at least $1 - \alpha$, $0 < \alpha < 1$, there holds,*

$$|R_N(w) - R(w)| \leq (b - a)\sqrt{\frac{1}{2N}\ln\left(\frac{2}{\alpha}\right)},$$

Here the probability measure is $\mathbb{P}_\mathcal{Z}^N$ on the space $(X \times Y)^N$ of N-tuples of samples S.

Proof. The N random variables $Z_i = \ell(x^{(i)}, y^{(i)}, w)$ satisfy the assumptions of Lemma 2.2 and there holds

$$\mu_i = \mu = \mathbb{E}_\mathcal{Z}(\ell(x, y, w)) = R(w).$$

Hence,

$$\mathbb{P}\{|R_N(w) - R(w)| > t\} = \mathbb{P}\{|\frac{1}{N}\sum_{i=1}^{N} Z_i - \mu| > t\} \leq 2\exp\left(\frac{-2Nt^2}{(b-a)^2}\right).$$

Solving the inequality

$$2\exp\left(\frac{-2Nt^2}{(b-a)^2}\right) \leq \alpha$$

for t we first obtain

$$\frac{-2Nt^2}{(b-a)^2} \leq \ln\left(\frac{\alpha}{2}\right)$$

and thus

$$t \geq (b-a)\sqrt{\frac{1}{2N}\ln\left(\frac{2}{\alpha}\right)}.$$

□

The bound in Theorem 2.3 is proportional to $b-a$, decreases like $1/\sqrt{N}$ as N increases and increases like $\ln(1/\alpha)$ as α decreases.

For estimating the term $\sup_{w\in\mathcal{W}}(R(w) - R_N(w))$ with high probability, we need a uniform version of Hoeffding's inequality. This requires to suitable measure the complexity of the set $\mathcal{W}$. More precisely, we need to characterize the complexity of the set of functions $\{\ell(\cdot,\cdot;w) \,:\, w \in W\}$.

Definition 2.4.

a) Let $\mathcal{U}$ be a set of indicator functions defined on $\mathcal{Z}$. The VC (Vapnik-Chervonenkis) dimension of $\mathcal{U}$ is defined as the largest number $l \in \mathbb{N}$ for which there exist points $z_1, \ldots, z_d \in \mathcal{Z}$ such that for any subset $J \subset \{1, \ldots, d\}$ (including the empty set) one can find $u \in \mathcal{U}$ with $u(z_j) = 1$ for all $j \in J$ and $u(z_j) = 0$ for all $j \notin J$.

b) Let $\mathcal{U}$ be a set of functions $u : Z \to \mathbb{R}$, satisfying $a \leq u(z) \leq b$ for all $z \in \mathcal{Z}$ and all $u \in \mathcal{U}$, where $-\infty < a < b < \infty$ are fixed. Then the VC dimension of $\mathcal{U}$ is defined as the VC dimension of the following set of indicator functions:

$$\{I[u(\cdot) - \beta \geq 0] \,:\, u \in \mathcal{U},\ \beta \in (a,b)\}.$$

Remark 2.5. There are also other concepts for defining measures for the complexity of a set of functions, e.g., the *Rademacher complexity*.

Example 2.6. For support vector machines, the VC-dimension of the set $\mathcal{U}$ of linear classifiers $I[v^T x + b \geq 0]$ is relevant, where $x \in \mathcal{X} = \mathbb{R}^n$, $v \in \mathbb{R}^n$ and $b \in \mathbb{R}$.

We show that the VC dimension is $d = n + 1$.

1. $d \geq n + 1$:

For the vectors $0, e_1, \ldots, e_n$ we achieve by choosing $b \in \{-1, 1\}$ and $v_i \in \{-2, 2\}$ that $I[v^T 0 + b \geq 0] = I[b \geq 0]$, $I[v^T e_i + b \geq 0] = I[v_i + b \geq 0] = I[v_i \geq 0]$. Thus,

we can achieve all 0-1 combinations by choosing b and v_i as above with the correct signs. This shows that the VC-dimension is at least $n+1$.

2. $d \leq n+1$:

Let $z_1, \ldots, z_{n+2} \in \mathbb{R}^n$ be arbitrarily fixed. The mapping

$$w = (v, b) \in \mathbb{R}^{n+1} \mapsto \begin{pmatrix} v^T z_1 + b \\ \vdots \\ v^T z_{n+2} + b \end{pmatrix} \in \mathbb{R}^{n+2} =: Mw$$

is linear and and the representing $(n+2) \times (n+1)$ matrix M with rows $(z_i^T, 1)$ has rank at most $n+1$. Thus, there exists $u \neq 0$ with $u^T M = 0$. By possibly changing the sign of u we achieve that $u_j < 0$ for some j. It is then not possible to find $w = (v, b)$ with

$$I[v^T z_i + b \geq 0] = I[u_i \geq 0] \quad \text{for all } i = 1, \ldots, n+2. \tag{8}$$

In fact, (8) would imply $u_i(v^T z_i + b) > 0$ if $u_i < 0$ and $u_i(v^T z_i + b) \geq 0$ if $u_i \geq 0$, which would lead to the following contradiction:

$$0 = u^T M w = \sum_i \underbrace{u_i(v^T z_i + b)}_{\geq 0} \geq u_j(v^T z_j + b) > 0.$$

Thus, we have found a 0-1 labeling of the $n+2$ points that cannot be achieved in the form $I[v^T z_i + b \geq 0]$, $1 \leq i \leq n+2$, by any choice of w. Since $z_1, \ldots, z_{n+2}$ were arbitrary, the VC dimension cannot satisfy $d \geq n+2$.

Using the VC-dismension, the following uniform estimate can be shown [Vap98, sec. 5.5]:

Theorem 2.7. *Let the family of loss functions $\ell(x, y, w)$, $w \in \mathcal{W}$ satisfy $0 \leq \ell(x, y, w) \leq b$ for all $(x, y, w) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{W}$ and have a finite VC-dimension d. Then with probability at least $1 - \alpha$, $0 < \alpha < 1$, and for $2N > d$, the following inequality holds for all $w \in \mathcal{W}$:*

$$R(w) - R_N(w) \leq \frac{b\eta}{2}\left(1 + \sqrt{1 + \frac{4R_N(w)}{b\eta}}\right) \leq \frac{b\eta}{2}\left(1 + \sqrt{1 + \frac{4}{\eta}}\right), \tag{9}$$

where

$$\eta = 4\frac{d\left(\ln \frac{2N}{d} + 1\right) - \ln(\alpha/4)}{N}.$$

The bound in (9) increases with b and with η. Further, η increases with d and also increases if α is decreased, and it decreases as N is increased.

With Theorems 2.3 and 2.7, we can upper bound the right hand side of (7). Targeting at a small such bound provides a guideline how to balance N, d, and ε.

3 Support Vector Machines

3.1 Classification by support vector machines

We consider the problem of binary classification, i.e., $|\mathcal{Y}| = 2$. In the following, we choose $\mathcal{Y} = \{-1, 1\}$ and assume that any input vector $x \in \mathcal{X} = \mathbb{R}^n$ ($\mathbb{R}^n$ this could be generalized to other Hilbert spaces) has a unique label $y = f(x)$, where $f : \mathcal{X} \to \mathcal{Y}$. The goal is to find within a suitable class of prediction functions $h(x; w)$ one which for all x predicts the corresponding label $y = f(x)$ with high reliability.

A natural choice for the loss function is then the 0-1-loss

$$\ell^{01}(x, y, w) = \begin{cases} 1 & \text{if } h(x; w) \neq y, \\ 0 & \text{if } h(x; w) = y. \end{cases}$$

For the 0-1-loss we obtain the expected risk

$$\begin{aligned} R^{01}(w) &= \int_{\mathcal{Z}} \ell^{01}(x, y, w)\, d\mathbb{P}_{\mathcal{Z}}(x, y) = \int_{\mathcal{X}} \ell^{01}(x, f(x), w)\, d\mathbb{P}_{\mathcal{X}}(x) \\ &= \mathbb{P}_{\mathcal{X}}(\{x \in \mathcal{X} \,:\, h(x; w) \neq f(x)\}). \end{aligned}$$

and the empirical risk

$$\begin{aligned} R_N^{01}(w) &= \frac{1}{N} \sum_{i=1}^{N} \ell^{01}(x^{(i)}, y^{(i)}, w) = \frac{1}{N} \left| \left\{ i \,:\, h(x^{(i)}) \neq y^{(i)} \right\} \right| \\ &= \frac{1}{N} \left| \left\{ i \,:\, h(x^{(i)}) \neq f(x^{(i)}) \right\} \right|. \end{aligned}$$

A disadvantage of this choice is that ℓ is nonsmooth and nonconvex w.r.t. w, which results in nonsmoothness and nonconvexity of the empirical risk. This is improved when one uses the hinge loss, which will arise in a natural way from the considerations that follow.

We still have to discuss how the class of prediction functions $h(\cdot; w)$ can be chosen.

Assuming $x \in \mathbb{R}^n$ (or x contained in a Hilbert space), a standard approach for classification is provided by support vector machines (SVM), where the idea is to separate the sets $\{x \in \mathcal{X} \,:\, f(x) = -1\}$ and $\{x \in \mathcal{X} \,:\, f(x) = 1\}$ as good as possible by a hyperplane $\{x \,:\, v^T x = b\}$. In this case, we have $w = (v, b) \in \mathbb{R}^{n+1}$. The idea then is to choose

$$h(x; w) = 2I[v^T x + b \geq 0] - 1 = \begin{cases} -1 & \text{if } v^T x + b < 0, \\ 1 & \text{if } v^T x + b \geq 0. \end{cases}$$

We first note that there holds $h(\cdot; tw) \equiv h(\cdot; w)$ for all w and all $t > 0$.

A parameter w^* that achieves a perfect classification, i.e., $\mathbb{P}_{\mathcal{X}}(\{x \,:\, h(x; w^*) = f(x)\} = 1$ (which is the same as $R^{01}(w^*) = 0$) is then only possible if $\{x \in \mathcal{X} \,:\, f(x) = -1\}$ and $\{x \in \mathcal{X} \,:\, f(x) = 1\}$ are (weakly) separable by a hyperplane for $\mathbb{P}_{\mathcal{X}}$-a.a. $x \in \mathcal{X}$.

Let us assume a bit more, namely that the two sets are strictly separable in the following sense:

Strict separability:

There exist v, b, and $\delta > 0$ such that

$$f(x)(v^T x + b) \geq \delta \quad \text{for } \mathbb{P}_{\mathcal{X}}\text{-a.a. } x \in \mathcal{X}. \tag{10}$$

By multiplying with $1/\delta$ we see that (10) is equivalent to (10)$|_{\delta=1}$ after replacing v and b with v/δ and b/δ:

There exist v and b such that

$$f(x)(v^T x + b) \geq 1 \quad \text{for } \mathbb{P}_{\mathcal{X}}\text{-a.a. } x \in \mathcal{X}. \tag{11}$$

Under strict separability there holds $h(x; w) = f(x)$ with probability 1 and thus $\ell^{01}(x, f(x), w) = 0$ with probability 1 which implies that w is a global minimizer of R^{01} with $R^{01}(w) = 0$.

We now show that in this strictly separable case, we can use the *hinge loss*

$$\ell^H(x, y, w) = \max(0, 1 - y(v^T x + b))$$

instead of the 0-1 loss to obtain an equivalent optimal classifier.

Lemma 3.1. *If $w = (v, b)$ satisfies the strict separability condition* (11)*, which also implies that w globally minimizes R^{01} with $R^{01}(w) = 0$, then for all $t \geq 1$, tw globally minimizes the hinge loss risk $R^H(w) = \mathbb{E}_{\mathcal{Z}}(\ell^H(\cdot, \cdot, w))$ with $R^H(tw) = 0$.*

Conversely, if w globally minimizes R^H with $R^H(w) = 0$, then (11) *holds and for all $t > 0$, tw globally minimizes R^{01} with $R^{01}(tw) = 0$.*

Proof. The condition (11) ensures $f(x)(v^T x + b) \geq 1$ for all $x \in M \subset \mathcal{X}$, where $\mathbb{P}_{\mathcal{X}}(M) = 1$. This implies $h(x; w) = f(x)$ for all $x \in M$. For all $t \geq 1$ and $x \in M$ we then have $f(x)(tv^T x + tb) \geq t \geq 1$, hence

$$\ell^H(x, f(x), tw) = \max(0, 1 - f(x)(tv^T x + tb)) \geq \max(0, 1 - t) = 0.$$

Thus, we get

$$R^H(tw) = \int_{\mathcal{X}} \ell^H(x, f(x), tw)\, d\mathbb{P}_{\mathcal{X}}(x) = \int_M \ell^H(x, f(x), tw)\, d\mathbb{P}_{\mathcal{X}}(x) = 0.$$

Conversely, $R^H(w) = 0$ implies $\ell^H(x, f(x), w) = 0$ on a set $M \subset \mathcal{X}$ with $\mathbb{P}_{\mathcal{X}}(M) = 1$. On M there then holds $f(x)(v^T x + b) \geq 1$. This shows that the strict separation property (11) holds for w. From this it follows that w is a global minimizer of R^{01} with $R^{01}(w) = 0$. The scale invariance $h(\cdot; tw) = h(\cdot; w)$ for all $t > 0$ shows $R^{01}(tw) = 0$ for all $t > 0$. □

The previous considerations can be easily transferred to empirical risk minimization:

Let $S = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ be an i.i.d. sample of length N.

The strict separability condition (11) then can be cast as follows:

There exist v, b such that

$$y^{(i)}(v^T x^{(i)} + b) \geq 1 \quad \text{for } i = 1, \ldots, N. \tag{12}$$

A straightforward adjustment of the previous arguments proves the empirical counterpart of Lemma 3.1:

Lemma 3.2. *If $w = (v, b)$ satisfies the strict separability condition* (12)*, which also implies that w globally minimizes R_N^{01} with $R_N^{01}(w) = 0$, then for all $t \geq 1$, tw globally minimizes the empirical hinge loss risk $R_N^H(w) = \mathbb{E}(\ell^H(\cdot, \cdot, w))$ with $R_N^H(tw) = 0$.*

Conversely, if w globally minimizes R_N^H with $R_N^H(w) = 0$, then (12) *holds and for all $t > 0$, tw globally minimizes R_N^{01} with $R_N^{01}(tw) = 0$.*

We now continue to focus on the empirical risk minimization problem and, for the case of strict separablity, investigate the question which of the many separating hyperplanes is the most robust choice.

The condition (12) compactly expresses the fact that

$$\begin{aligned} v^T x^{(i)} &\leq -b - 1 \quad \text{for all } i \text{ with } y^{(i)} = -1, \\ v^T x^{(i)} &\geq -b + 1 \quad \text{for all } i \text{ with } y^{(i)} = 1. \end{aligned}$$

The distance between these hyperplanes $v^T x = b \pm 1$ is given by $2/\|v\|$ and in the strip between them there does not lie any sample point $x^{(i)}$. The hyperplane $v^T x = b$ lies in the middle between these two.

A natural idea for finding a robust separating hyperplane is to choose $w = (v, b)$ such that (12) holds and the distance $2/\|v\|$ is maximized.

This can be written as

$$\begin{aligned} \max_{v\ (\neq 0), b} \quad & \frac{2}{\|v\|} \\ \text{s.t.} \quad & y^{(i)}(v^T x^{(i)} + b) \geq 1 \quad (i = 1, \ldots, N). \end{aligned}$$

We can rewrite this as a quadratic program

$$\begin{aligned} \min_{v,b} \quad & \frac{1}{2}\|v\|^2 \\ \text{s.t.} \quad & y^{(i)}(v^T x^{(i)} + b) \geq 1 \quad (i = 1, \ldots, N). \end{aligned}$$

This formulation can provide a basis for computing $w = (v, b)$.

If separation is not possible, then we can introduce a *soft margin* by relaxing the constraints to

$$y^{(i)}(v^T x^{(i)} + b) + z_i \geq 1, \quad z_i \geq 0 \quad (i = 1, \ldots, N),$$

and adding a penalty to the cost function. We obtain

$$\begin{aligned} \min_{v,b,z} \quad & \frac{e^T z}{N} + \frac{\gamma}{2}\|v\|^2 \\ \text{s.t.} \quad & y^{(i)}(v^T x^{(i)} + b) + z_i \geq 1, \quad z_i \geq 0 \quad (i = 1, \ldots, N). \end{aligned} \tag{13}$$

As we will now show, this problem can be written as a regularized empirical risk minimization problem using the hinge loss as follows:

$$\min_{w=(v,b)} \quad R_N^H(w) + \frac{\gamma}{2}\|v\|_2^2, \tag{14}$$

where $R^H(w) = \frac{1}{N}\sum_{i=1}^N \max(0, 1 - y^{(i)}(v^T x^{(i)} + b))$. This can be rewritten as

$$\begin{aligned} \min_{v,b,z} \quad & \frac{1}{N}\sum_{i=1}^N z_i + \frac{\gamma}{2}\|v\|^2 \\ \text{s.t.} \quad & 1 - y^{(i)}(v^T x^{(i)} + b) \leq z_i, \quad z_i \geq 0 \quad (i = 1, \ldots, N). \end{aligned}$$

and this is the same problem as (13).

The parameter γ in (13) or (14) can be used to control the richness of the set of vectors $v \in \mathbb{R}^n$ that are relevant when minimizing the (empirical) risk. It thus implicitly can help to reduce the VC-dimension.

One can show (exercises) that there exists a solution to (13) with $v_N^* \neq 0$ if there exist i and i' with $y^{(i)}y^{(i')} = -1$. Further, it can be shown that in the strictly separable case, there holds $z^* = 0$ if $\gamma > 0$ is sufficiently small.

There are also variants of SVM where a different regularization term is used, e.g., the ℓ_1-norm:

$$\min_{w=(v,b)} \quad R_N^H(w) + \gamma\|v\|_1.$$

One can show that this penalty term promotes 0-entries in the vector v of optimal solutions. Therefore, the effective dimension of vectors v is reduced, which reduces the VC dimension.

3.2 KKT conditions and dual problem for the SVM QP

The problem (13) is convex with linear constraints and thus the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient. For deriving them, we introduce the Lagrange function

$$L(v, b, z, \lambda, \mu) = \frac{e^T z}{N} + \frac{\gamma}{2}\|v\|^2 - \sum_{i=1}^N \lambda_i(y^{(i)}(v^T x^{(i)} + b) + z_i - 1)) - \mu^T z.$$

Now, by the KKT theorem (the constraints are linear, which is a constraint qualification, and the problem is convex), $w_N^* = (v_N^*, b_N^*)$ and z^* solve (13) iff there exist

vectors of Lagrange multipliers $\lambda^*, \mu^* \in \mathbb{R}^N$ such that:
(15)

$$\nabla_v L(v_N^*, b_N^*, z^*, \lambda^*, \mu^*) = \gamma v_N^* - \sum_{i=1}^N \lambda_i^* y^{(i)} x^{(i)} = 0,$$

$$\nabla_b L(v_N^*, b_N^*, z^*, \lambda^*, \mu^*) = -\sum_{i=1}^N \lambda_i^* y^{(i)} = 0,$$

$$\nabla_z L(v_N^*, b_N^*, z^*, \lambda^*, \mu^*) = \frac{1}{N} e - \lambda^* - \mu^* = 0,$$

For $i = 1, \ldots, N$:

$$y^{(i)}((v_N^*)^T x^{(i)} + b_N^*) + z_i^* \geq 1, \quad \lambda_i^* \geq 0, \quad (y^{(i)}((v_N^*)^T x^{(i)} + b_N^*) + z_i^* - 1)\, \lambda_i^* = 0,$$
$$z_i^* \geq 0, \quad \mu_i^* \geq 0, \quad z_i^* \mu_i^* = 0.$$

It is well known that the Lagrange multipliers are the solution of the Lagrangian dual problem that we will derive below. Further, the KKT condition can also be interpreted as the KKT conditions of the dual problem, where the primal variables (v, b, z) play the role of the Lagrange multipliers of the dual problem.

From the first KKT condition it becomes apparent that if the multiplier vector $\lambda^* \in \mathbb{R}^N$ is known, then the normal vector of the optimal hyperplane can be recovered via

$$v_N^* = \frac{1}{\gamma} \sum_{i=1}^N \lambda_i^* y^{(i)} x^{(i)}.$$

In this sum, only those i with $\lambda_i^* > 0$ make a relevant contribution. The corresponding vectors $x^{(i)}$ are called *support vectors*. For $\lambda_i^* > 0$ it is required that the corresponding constraint is active, i.e., that the corresponding soft margin constraint holds with equality:

$$y^{(i)}((v_N^*)^T x^{(i)} + b_N^*) + z_i^* = 1.$$

In the strictly separable case and if $z^* = 0$, this means that the support vectors lie on the hyperplane

$$y^{(i)}(v_N^*)^T x^{(i)} = 1 - y^{(i)} b_N^*.$$

This is the closer one of the two hyperplanes that separate the two classes of points with maximum margin.

Based on this, in view of solving the problem (13) via the dual problem, we now discuss how the primal solution v_N^*, b_N^*, z^* can be recovered from the multipliers λ^* and μ^*. We know this already for v_N^*.

In the case that there exists i with $\lambda_i^* > 0$ and $\mu_i^* > 0$ we get

$$z_i^* = 0 \quad \text{and} \quad y^{(i)}((v_N^*)^T x^{(i)} + b_N^*) + z_i^* = 1,$$

hence, using $1/y^{(i)} = y^{(i)}$:

$$z_i^* = 0, \quad b_N^* = y^{(i)} - (v_N^*)^T x^{(i)}.$$

It remains to consider the case where for all i either $\lambda_i^* = 1/N$ and $\mu_i^* = 0$ or $\lambda_i^* = 0$ and $\mu_i^* = 1/N$. For all i with $\lambda_i^* = 1/N$ we then have

$$y^{(i)}((v_N^*)^T x^{(i)} + b_N^*) + z_i^* = 1, \quad z_i^* \geq 0$$

and for all i with $\lambda_i^* = 0$ we have

$$z_i^* = 0, \quad y^{(i)}((v_N^*)^T x^{(i)} + b_N^*) \geq 1.$$

Any choice of z^* and b_N^* satisfying these conditions is suitable. Viewing z_i^* as a slack variable, we obtain a set of lower and upper bounds for b_N^*. We then can choose b_N^* lying between the maximum lower bound and the minimum upper bound. Finally, the z_i^* can be set appropriately.

We now derive the Lagrange dual problem: The dual objective function is defined as

$$\begin{aligned} d(\lambda, \mu) &= \inf_{v,b,z} \; L(v, b, z, \lambda, \mu) \\ &= \sum_{i=1}^N \lambda_i + \inf_v \frac{\gamma}{2}\|v\|^2 - \sum_{i=1}^N \lambda_i y^{(i)} v^T x^{(i)} \\ &+ \inf_b -b \sum_{i=1}^N \lambda_i y^{(i)} + \inf_z \frac{e^T z}{N} - \sum_{i=1}^N (\lambda_i + \mu_i) z_i. \end{aligned}$$

The minimium w.r.t. v is attained iff

$$(\nabla_v L(v, b, z, \lambda, \mu) =) \; \gamma v - \sum_{i=1}^N \lambda_i y^{(i)} x^{(i)} = 0$$

and the resulting minimal value is

$$-\frac{1}{2\gamma} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)}.$$

Under $\inf_b$ and $\inf_z$ there are linear functions. This gives

$$d(\lambda, \mu) = \sum_{i=1}^N \lambda_i - \frac{1}{2\gamma} \sum_{i=1}^N \sum_{j=1}^N \lambda_i \lambda_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)}$$

if

$$\begin{aligned} (\nabla_b L(v, b, z, \lambda, \mu) =) \; & -\sum_{i=1}^N \lambda_i y^{(i)} = 0, \\ (\nabla_z L(v, b, z, \lambda, \mu) =) \; & \frac{e}{N} - \lambda - \mu = 0, \end{aligned} \tag{16}$$

and $d(\lambda, \mu) = -\infty$, otherwise.

Note that the conditions in (16) are exactly the 2nd and 3rd equations in the KKT conditions (15). They form the constraints of the dual problem:

$$\begin{aligned} \max_{\lambda,\mu} \; & e^T\lambda - \tfrac{1}{2}\lambda^T K\lambda \\ \text{s.t.} \; & \sum_{i=1}^N \lambda_i y^{(i)} = 0, \\ & \frac{e}{N} - \lambda - \mu = 0, \\ & \lambda \geq 0, \quad \mu \geq 0, \end{aligned} \tag{17}$$

where

$$K = (k_{ij}) \in \mathbb{R}^{N\times N}, \quad k_{ij} = \tfrac{1}{\gamma} y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)}.$$

The multiplier μ can be removed via the second constraint: $\mu = e/N - \lambda$. We then get

$$\begin{aligned} \max_{\lambda} \; & e^T\lambda - \tfrac{1}{2}\lambda^T K\lambda \\ \text{s.t.} \; & \sum_{i=1}^N \lambda_i y^{(i)} = 0, \\ & 0 \leq \lambda \leq \frac{1}{N} e. \end{aligned} \tag{18}$$

3.3 The kernel trick

If in a classification problem the two classes of input data corresponding to the labels -1 and 1 cannot approximately be separated by a hyperplane, then the SVM is not directly applicable. It is then however often possible, to first map the data $x \in \mathcal{X}$ to corresponding features $\tilde{x} \in \tilde{\mathcal{X}}$, where the feature space $\tilde{\mathcal{X}}$ is a Hilbert space with inner product $(\cdot,\cdot)_{\tilde{\mathcal{X}}}$. For instance, $\tilde{\mathcal{X}}$ could be a $\mathbb{R}^{\tilde{n}}$. Denote the operator used to map x to $\tilde{x}$ by $\Phi : \mathcal{X} \to \tilde{\mathcal{X}}$. We now can apply an SVM to the features $\tilde{x}^{(i)} = \Phi(x^{(i)})$, which are labeled by $y^{(i)}$.

We thus have to replace every occurance of x or $x^{(i)}$ by $\Phi(x)$ or $\Phi(x^{(i)})$ and we have to use the inner product on $\tilde{\mathcal{X}}$. Doing so, the prediction function is

$$h(x;\tilde{w}) = 2I[(\tilde{v}, \Phi(x))_{\tilde{\mathcal{X}}} + \tilde{b} \geq 0] - 1,$$

where $\tilde{v} \in \tilde{X}$ and $\tilde{b}$. We will now see that it is not needed to compute or store feature vectors explicitly.

In fact, transcribing the QP of the SVM, we get

$$\begin{aligned} \min_{\tilde{v},\tilde{b},z} \; & \frac{e^T z}{N} + \frac{\gamma}{2}\|\tilde{v}\|^2_{\tilde{\mathcal{X}}} \\ \text{s.t.} \; & y^{(i)}((\tilde{v}, \Phi(x^{(i)}))_{\tilde{\mathcal{X}}} + \tilde{b}) + z_i \geq 1 \quad i = 1,\dots,N, \end{aligned}$$

and for the dual problem, we obtain

$$\max_{\lambda\in\mathbb{R}^N} \; e^T\lambda - \tfrac{1}{2}\lambda^T K\lambda \quad \text{s.t.} \quad \sum_{i=1}^N \lambda_i y^{(i)} = 0, \quad 0 \leq \lambda \leq \tfrac{1}{N} e.$$

Here,

$$K = (k_{ij}), \quad \text{where} \quad k_{ij} = \frac{1}{\gamma} y^{(i)} y^{(j)} (\Phi(x^{(i)}), \Phi(x^{(j)}))_{\tilde{\mathcal{X}}}.$$

We introduce the *kernel function*

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad k(x, x') = (\Phi(x), \Phi(x'))_{\tilde{\mathcal{X}}}.$$

Then we have

$$K = \frac{1}{\gamma} y^{(i)} y^{(j)} k(x^{(i)}, x^{(j)}).$$

Using the solution λ^* of the dual problem, we get as before from the KKT conditions that

$$\tilde{v}_N^* = \gamma \sum_{\lambda_i^* > 0} \lambda_i^* y^{(i)} \Phi(x^{(i)}).$$

Inserting this into the prediction function, we obtain

$$\begin{aligned} h(x; \tilde{w}) &= 2I[(\tilde{v}_N^*, \Phi(x))_{\tilde{\mathcal{X}}} + \tilde{b} \geq 0] - 1 \\ &= 2I\left[\gamma \sum_{\lambda_i^* > 0} \lambda_i^* y^{(i)} (\Phi(x^{(i)}, \Phi(x))_{\tilde{\mathcal{X}}} + \tilde{b} \geq 0\right] - 1 \\ &= 2I\left[\gamma \sum_{\lambda_i^* > 0} \lambda_i^* y^{(i)} k(x^{(i)}, x) + \tilde{b} \geq 0\right] - 1. \end{aligned}$$

Thus, for formulating the dual problem and for expressing the prediction function, we only need the kernel function k, not the function Φ.

Often, one does not explicitly choose Φ, but rather one chooses the kernel k. Under suitable assumptions, one can show that for a given kernel there exist a corresponding mapping Φ:

Theorem 3.3 (Mercer's Theorem)**.** *Let $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a symmetric positive semidefinite kernel, i.e.:*

a) *k is continuous.*

b) *k is symmetric, i.e., $k(x, x') = k(x', x)$ for all $x, x' \in \mathbb{R}^n$.*

c) *k is positive semidefinite in the sense that for every finite collection of points $x_1, \ldots, x_m \in \mathbb{R}^n$, $m \in \mathbb{N}$, the $(m \times m)$-matrix with (i, j)-entry $k(x_i, x_j)$ is positive semidefinite.*

Further, let μ be a strictly positive Borel measure on $\mathbb{R}^n$ (e.g., the Lebesgue measure) with

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^n} k(x, x')^2 \, d\mu(x) d\mu(x') < \infty$$

Then there exists a sequence $(\lambda_l) \subset \mathbb{R}_+$ and an orthonormal basis $(\psi_l) \subset L^2_\mu(\mathbb{R}^d)$ such that

$$k(x, x') = \sum_{l=1}^{\infty} \lambda_l \psi_l(x) \psi_l(x') \quad \forall \, x, x' \in \mathbb{R}^n.$$

The sequence converges absolutely for every $(x, x') \in \mathbb{R}^n \times \mathbb{R}^n$ and uniformly on compact subsets of $\mathbb{R}^n \times \mathbb{R}^n$.

Given such a kernel k, we thus can write it as

$$k(x, x') = (\Phi(x), \Phi(x'))_{\tilde{\mathcal{X}}},$$

where

$$\Phi(x) = (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots) \in \ell_2, \quad (\tilde{x}, \tilde{x}')_{\tilde{X}X} = (\tilde{x}, \tilde{x}')_{\ell_2} = \sum_{l=1}^{\infty} \tilde{x}_l \tilde{x}'_l.$$

Here are several kernels that are often used in practice:

- $k(x, x') = x^T x'$ (linear),
- $k(x, x') = \exp(-\theta\|x - x'\|^2)$ (Gaussian),
- $k(x, x') = (x^T x' + 1)^q$ (polynomial).

3.4 Numerical solution of the dual SVM problem

The dual formulation in the form (18) is a QP that has a dimension independent of the dimension of the input space $\mathcal{X}$, can be used also with kernels, and has relatively simple constraints: Bounds on the variables and a single equality constraint. It thus provides a good basis for numerical solution methods. One possibility is to apply efficient implementations of QP solvers (e.g., interior-point or active set methods) to the dual problem. The problems, however, can be very large and it is desirable not to form and store the complete matrix K but only to access small parts of K to compute each step.

Thus, researchers have looked for methods that are particularly well adjusted to the specific problem characteristics. An idea to achieve that only small parts of K need be computed and stored at a time is to generate a sequence of feasible iterates λ^k and to update only a few variables λ_i, $i \in B_k$, at each iteration k while keeping the others at their current value λ_i^k. Here, λ^k denotes the iterate at the begin of iteration k. For updating the dual problem we thus choose a small subset (working set) $B_k \subset \{1, \ldots, N\}$, $|B_k| = q$ and solve the dual problem restricted to λ_{B_k}, while $\lambda_{B_k^c} = \lambda^k_{B_k^c}$ is fixed, where $B_k^c = \{1, \ldots, N\} \setminus B_k$.

We will now develop and analyze this approach, called *decomposition method* (a terminology that I don't find quite appropriate; better would be subspace optimization method or overlapping block coordinate descent).

We consider QPs of the more general form:

$$\min_{\lambda} \; f(\lambda) := c^T\lambda + \frac{1}{2}\lambda^T K\lambda \quad \text{s.t.} \quad A\lambda = b, \quad 0 \le \lambda \le a, \tag{19}$$

where $A \in \mathbb{R}^{m\times N}$, $b \in \mathbb{R}^m$, $a \in \mathbb{R}^N_{++}$, and $K = K^T \in \mathbb{R}^{N\times N}$ is positive semidefinite.

We introduce the feasible set $\Lambda = \{\lambda \in \mathbb{R}^N \,:\, A\lambda = b,\ 0 \le \lambda \le a\}$.

For the dual SVM problem (18) we have (note that it is a maximization problem which we have to convert to a minimization problem) $c = -e$, $A = (y^{(1)}, \ldots, y^{(N)})$, $b = 0$, $a = \frac{1}{N}e$, $m = 1$.

Let us denote by $\mathcal{B}$ be the set of all subsets $B \subset \{1, \ldots, N\}$ with $|B| = q$, where $q \in \{m+1, \ldots, N\}$ is fixed. The sets B are called working sets.

New iterates are computed based on

$$\text{QP}(B, \hat{\lambda}_{B^c}) \qquad \min_{\lambda_B} f(\lambda|_{\lambda_{B^c}=\hat{\lambda}_{B^c}}) \quad \text{s.t.} \quad \lambda \in \Lambda.$$

where $B \in \mathcal{B}$ and $B^c = \{1, \ldots, N\} \setminus B$. Here, $\hat{\lambda}$ is the current iterate.

This can equivalent be written as (omitting constant offsets in the cost function):

$$\text{QP}(B, \hat{\lambda}_{B^c}) \qquad \begin{aligned} &\min_{\lambda_B} \ (c_B + K_{BB^c}\hat{\lambda}_{B^c})^T\lambda_B + \tfrac{1}{2}\lambda_B^T K_{BB}\lambda_B \\ &\text{s.t. } A_B\lambda_B = b_B - A_B\hat{\lambda}_{B^c}, \quad 0 \le \lambda_B \le a_B. \end{aligned}$$

As proposed in [LS04], we build the algorithm upon a family of criticality measures $C_B(\lambda)$, $B \in \mathcal{B}$.

Assumption 3.4.

1. For all $B \in \mathcal{B}$, the function $C_B : \Lambda \to \mathbb{R}_+$ is continuous.
2. For any $\lambda \in \Lambda$ and any $B \in \mathcal{B}$, there holds $C_B(\lambda) = 0$ iff λ_B solves $\text{QP}(B, \lambda_{B^c})$.
3. If λ is not optimal for (19) then there exists a working set $B \subset \{1, \ldots, N\}$, $|B| = q$, such that $C_B(\lambda) > 0$.

We will discuss later how a suitable family of functions C_B can be defined.

We consider the following optimization method for solving (19):

Algorithm 3.5 (Decomposition method)**.**

0. Choose $\lambda^0 \in \Lambda$.

For $k = 0, 1, \ldots$:

1. Choose a working set $B_k \in \mathcal{B}$ such that $C_{B_k}(\lambda^k)$ achieves a fraction of the maximum possible value over all working sets:
$$C_{B_k}(\lambda^k) \ge \eta \max_{B \in \mathcal{B}} C_B(\lambda^k).$$
2. If $C_{B_k}(\lambda^k) = 0$, then STOP (λ^k solves (19)).
3. Compute $\lambda^{k+1}_{B_k}$ by solving $\text{QP}(B_k, \lambda_{B_k^c})$ and set $\lambda^{k+1}_{B_k^c} = \lambda^k_{B_k^c}$.

We first show a result that relates $\|\lambda^k - \lambda^{k+1}\|$ to the cost function decrease $f(\lambda^k) - f(\lambda^{k+1})$:

Lemma 3.6. *Assume that there exists $\sigma > 0$ such that there holds*

$$\lambda_{\min}(K_{BB}) \geq \sigma > 0 \quad forall\ B \in \mathcal{B}. \tag{20}$$

Then for the sequence generated by Algorithm 3.5 there holds

$$f(\lambda^k) - f(\lambda^{k+1}) \geq \frac{\sigma}{2}\|\lambda^{k+1} - \lambda^k\|^2.$$

Proof. Since $\lambda^{k+1}_{B_k}$ is a solution of QP$(B_k, \lambda^k_{B_k^c})$ and $\lambda^k_{B_k}$ is feasible for this problem, we obtain from the optimality conditions and Taylor's theorem:

$$\begin{aligned} 0 &\leq \nabla_{\lambda_{B_k}} f(\lambda^{k+1})^T(\lambda^k_{B_k} - \lambda^{k+1}_{B_k}) \\ &= f(\lambda^k) - f(\lambda^{k+1}) - \frac{1}{2}(\lambda^k_{B_k} - \lambda^{k+1}_{B_k})^T K_{B_k B_k}(\lambda^k_{B_k} - \lambda^{k+1}_{B_k}). \end{aligned}$$

This yields

$$\begin{aligned} f(\lambda^k) - f(\lambda^{k+1}) &\geq \frac{1}{2}(\lambda^k_{B_k} - \lambda^{k+1}_{B_k})^T K_{B_k B_k}(\lambda^k_{B_k} - \lambda^{k+1}_{B_k}) \\ &\geq \frac{\sigma}{2}\|\lambda^{k+1}_{B_k} - \lambda^k_{B_k}\|^2 = \frac{\sigma}{2}\|\lambda^{k+1} - \lambda^k\|^2. \end{aligned}$$

□

Theorem 3.7. *Let Assumption* 3.4 *hold and assume that there exists $\sigma > 0$ such that* (20) *holds. Let the sequence (λ^k) be generated by Algorithm* 3.5*. Then every accumulation point $\bar{\lambda}$ of (λ^k) is a solution of* (19)*.*

Proof. Let $\bar{\lambda}$ be an accumulation point of (λ^k) and denote by $(\lambda^k)_K$ a subsequence converging to $\bar{\lambda}$. From the closedness of Λ and $\lambda^k \in \Lambda$ we conclude $\bar{\lambda} \in \Lambda$. Now assume that $\bar{\lambda}$ is not optimal for (19). Then by Assumption 3.4, there exist a working set $\bar{B}$ and $\varepsilon > 0$ with

$$C_{\bar{B}}(\bar{\lambda}) \geq 3\varepsilon > 0.$$

By continuity of every function in the finite family C_B, $B \in \mathcal{B}$, there exist $\delta > 0$ with

$$|C_B(\lambda) - C_B(\bar{\lambda})| < \eta\varepsilon \quad \forall \lambda \in B_\delta(\bar{\lambda}) \cap \Lambda,\ \forall\ B \in \mathcal{B}.$$

Since the algorithm is a descent method, $(f(\lambda^k))$ decreases. Together with $(f(\lambda^k))_K \to f(\bar{\lambda})$, this yields $f(\lambda^k) \downarrow f(\bar{\lambda})$. We thus obtain $f(\lambda^{k+1}) - f(\lambda^k) \to 0$ as $k \to \infty$.

Hence, there holds for all $k \in K$:

$$\begin{aligned} \|\lambda^{k+1} - \bar{\lambda}\| &\leq \|\lambda^{k+1} - \lambda^k\| + \|\lambda^k - \bar{\lambda}\| \\ &\leq \sqrt{\frac{2}{\sigma}(f(\lambda^k) - f(\lambda^{k+1}))} + \|\lambda^k - \bar{\lambda}\| \to 0 \quad \text{as } K \ni k \to \infty. \end{aligned}$$

Therefore, there exists $l \geq 0$ with

$$\lambda^k, \lambda^{k+1} \in B_\delta(\bar{\lambda}) \cap \Lambda \quad \forall\ k \in K,\ k \geq l.$$

Hence, we have for all $k \in K$, $k \geq l$:

$$C_{\bar{B}}(\lambda^k) > C_{\bar{B}}(\bar{\lambda}) - \eta\varepsilon \geq (3-\eta)\varepsilon.$$

Using

$$C_{B_k}(\lambda^k) \geq \eta C_{\bar{B}}(\lambda^k) > (3-\eta)\eta\varepsilon \quad \forall k \in K,\ k \geq l,$$

we get

$$C_{B_k}(\lambda^{k+1}) > C_{B_k}(\bar{\lambda}) - \eta\varepsilon > C_{B_k}(\lambda^k) - 2\eta\varepsilon > (1-\eta)\eta\varepsilon \geq 0.$$

This contradicts $C_{B_k}(\lambda^{k+1}) = 0$, which holds since λ_B^{k+1} solves QP$(B_k, \lambda^k_{B_k^c})$.

Therefore, the assumption that $\bar{\lambda}$ does not solve (19) was wrong. □

We next show that Assumption 3.4 is satisfied if we choose $C_B : \Lambda \mapsto \mathbb{R}$ as the *gain*

$$C_B(\lambda) = f(\lambda) - f(\lambda^*(B,\lambda)), \tag{21}$$

where $[\lambda^*(B,\lambda)]_B \in \mathbb{R}^q$ is the optimal solution of QP(B, λ_{B^c}), and $[\lambda^*(B,\lambda)]_{B^c} = \lambda_{B^c}$.

Note that $f(\lambda^*(B,\lambda))$ is the optimal function value of QP(B, λ_{B^c}).

It is clear that this choice of C_B satisfies Assumption 3.4, 2.

The following Lemma shows that the gain also satisfies Assumption 3.4, 3.

Lemma 3.8. *If $\lambda \in \mathbb{R}^N$ is not optimal for* (19) *then there exists $B \in \mathcal{B}$ such that λ_B is not optimal for* QP(B, λ_{B^c}). *In particular, $C_B(\lambda) > 0$, where C_B is defined as in* (21).

Proof. If λ is not optimal for (19), then there exists $d \neq 0$ with $Ad = 0$, $\lambda + d \in [0,a]$, and $\nabla f(\lambda)^T d < 0$. We set $J_0 = \{i : \lambda_i = 0\}$, $J_a = \{i : \lambda_i = a_i\}$ and $J = J_0 \cup J_a$. Then the following problem has a nonzero solution with a negative optimal value:

$$\min_d \nabla f(\lambda)^T d \quad \text{s.t.} \quad d|_{J_0} \geq 0, \quad d|_{J_a} \leq 0, \quad Ad = 0, \quad \|d\|_1 = 1. \tag{22}$$

We write this as an LP:

$$\begin{aligned} \min_{d, v_{J^c}} \ & \nabla f(\lambda)^T d \\ \text{s.t.} \ & Ad = 0, \quad e^T d_{J_0} - e^T d_{J_a} + e^T v_{J^c} = 1, \\ & d_{J_0} \geq 0, \quad d_{J_a} \leq 0, \quad -v_{J^c} \leq d_{J^c} \leq v_{J^c}. \end{aligned} \tag{23}$$

We use the notation v_{J^c} to indicate that v_{J^c} is a vector with $|J^c| = n - |J|$ components v_i, $i \in J^c$.

The problem has $2N - |J|$ variables and the gradient of the constraints span $\mathbb{R}^{2N-|J|}$. In fact, already the gradients of the $|J| + 2|J^c| = 2N - |J|$ constraints

(i) $-d|_{J_0} \leq 0$, (ii) $d|_{J_a} \leq 0$, (iii) $-d_{J^c} - v_{J^c} \leq 0$, (iv) $d_{J^c} - v_{J^c} \leq 0$

form a matrix of rank $2N - |J|$.

By construction, the LP (23) has a solution and due to the full rank of the constraint matrix, there exist a solution that is an optimal vertex $\binom{d^*}{v^*_{Jc}}$. At this solution the gradients of active constraints span the full space $\mathbb{R}^{2N-|J|}$. Thus, at this solution at least $2N - |J|$ constraints are active. Hence, at least $2N - |J| - m - 1$ of the $2N - |J|$ constraints in (i)–(iv) are active at the optimal vertex.

It is now not very difficult to see that this implies $d^*_i = 0$ for at least $N - m - 1$ different indices $i \in \{1, \dots, N\}$.

We thus have proved that d^* has at most $m + 1 \leq q$ nonzero components. Hence, we can find $B \in \mathcal{B}$ such that $d^*_i = 0$ for all $i \notin B$. We noted before that the optimal value of (22) is negative, hence $\nabla_{\lambda_B} f(\lambda)^T d^*_B = \nabla f(\lambda)^T d^* < 0$.

The constructed d^* satisfies:

$$\nabla_{\lambda_B} f(\lambda)^T d^*_B < 0, \quad d^*_{B^c} = 0, \quad A_{\cdot,B} d^*_B = 0, \quad d^*_{B \cap J_0} \geq 0, \quad d^*_{B \cap J_a} \leq 0.$$

This shows that λ_B is not optimal for QP(B, λ_{B^c}) and thus we have found $B \in \mathcal{B}$ with $C_B(\lambda) > 0$. □

It remains to prove that C_B as defined in (21) is continuous on Λ. This can be done by using results on the value function for parametric QPs. For the concrete case of (18), a direct proof of the continuity of the gain C_B is given in [GI06,Lemma 5].

Hence, the gain defined in (21) satisfies Assumption 3.4.

Algoritm 3.5 is thus convergent if we use the gain as criticality measure.

We now return to the special problem (18), which we write in the form (19) with $c = -e$, $A = (y^{(1)}, \dots, y^{(N)})$, $b = 0$, $a = \frac{1}{N} e$, $m = 1$, and we consider the working set size $q = 2 = m + 1$.

The solution of the subproblem QP(B, λ_{B^c}) can then be reduced to a 1-dimensional problem and only the i-th and j-th rows of K are needed, where $B = \{i, j\}$. Details are given in the exercises.

To reduce the number of working sets that have to be considered, Glasmachers and Igel [GI06] propose a hybrid approach where, whenever conditions are satisfied that preserve convergence, the maximum gain working set B_k, $k \geq 2$, is chosen among all the $2N - 4$ working sets in $\mathcal{B}$ that satisfy $|B_k \cap B_{k-1}| = 1$. In cases where this choice of the working set is not good enough, they resort to a more expensive choice. They use the *most violating pair* (MVP) strategy, see the exercises, but one also could use the maximum gain strategy, i.e., like in step 2 of Alg. 3.5 with C_B given by (21).

Decomposition methods for SVM are also called SMO (*sequential minimal optimization*) methods.

4 Stochastic Gradient Methods

We now consider methods for solving

$$\min_w \ f(w) \quad \text{s.t.} \quad w \in \mathcal{W}, \tag{24}$$

where $\mathcal{W} \subset \mathbb{R}^d$ is closed and convex and $f : \mathcal{U} \to \mathbb{R}$ is convex on an open convex neighborhood $\mathcal{U}$ of $\mathcal{W}$.

We allow for nonsmoothness of the function f, since, e.g., the hinge loss (as used for SVM) is convex, but not everywhere differentiable.

In the situations that we have in mind here, there usually holds either

$$f(w) = \mathbb{E}_{\mathcal{Z}} F(w, \cdot), \tag{25}$$

or

$$f(w) = \frac{1}{N} \sum_{i=1}^{N} F(w, z^{(i)}), \quad z^{(i)} \in \mathcal{Z}. \tag{26}$$

As noted earlier, the average can be interpreted as an expectation as well, with a discrete probability distribution on the samples.

Thus, if F is the loss, then in the two cases f coincides with the risk and the empirical risk, respectively.

We require that $F : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ is such that

$$w \mapsto F(w, z)$$

is convex on the open convex neighborhood $\mathcal{U}$ of $\mathcal{W}$ for all $z \in \mathcal{Z}$ and

$$z \in \mathcal{Z} \mapsto F(w, z)$$

is integrable for all $w \in \mathcal{U}$.

Then $f : \mathcal{W} \to \mathbb{R}$ is convex on $\mathcal{U}$. In fact, using $\mathbb{E}_{\mathcal{Z}} v_1 \leq \mathbb{E}_{\mathcal{Z}} v_2$ if $v_1(z) \leq v_2(z)$ for $\mathbb{P}_{\mathcal{Z}}$-a.a. $z \in \mathcal{Z}$ we get for all $w_1, w_2 \in \mathcal{U}$, $0 \leq t \leq 1$:

$$\mathbb{E}_{\mathcal{Z}} F((1-t)w_1 + tw_2, \cdot) \leq \mathbb{E}_{\mathcal{Z}}[(1-t)F(w_1, \cdot) + tF(w_2, \cdot)] = (1-t)f(w_1) + tf(w_2).$$

More generally than in the situations (25) and (26), *stochastic subgradient methods*, which we now will develop, are applicable to any problem of the form (24) as long as the setting given in Assumption 4.2 below is met.

4.1 Stochastic subgradient method

In the deterministic case, the *subgradient method* for solving (24) chooses

$$w^{k+1} = \pi_{\mathcal{W}}(w^k - \eta_k g^k)$$

as the new iterate, where $g^k \in \partial f(w^k)$ is a subgradient, $\eta_k > 0$ is a step size and $\pi_{\mathcal{W}}$ is the metric projection onto $\mathcal{W}$.

Here, the *subdifferential* $\partial f(w)$, $w \in \mathcal{U}$, consists of all vectors $g \in \mathbb{R}^n$, called *subgradients*, such that

$$f(\hat{w}) - f(w) \geq g^T(\hat{w} - w) \quad \forall\, \hat{w}, w \in \mathcal{U}.$$

We will later need the following fact from convex analysis:

$$\forall\, w \in \mathcal{U},\ s \in \mathbb{R}^n \quad \exists\, g \in \partial f(w): \quad f'(w,s) = g^T s.$$

Here,

$$f'(w,s) = \lim_{t \to 0^+} \frac{f(w+ts) - f(w)}{t}$$

is the directional derivative.

In stochastic subgradient methods, we use a stochastic approximation $\hat{g}^k \in \mathbb{R}^d$ to an element of $\partial f(w^k)$.

Assumption 4.1. The following setting is given:

a) A probability space $(\Xi, \mathbb{P}_\Xi)$.

b) A method to draw a sequence of independent samples $\xi^1, \xi^2, \ldots \in \Xi$ from $(\Xi, \mathbb{P}_\Xi)$.

c) A function $G : \mathcal{U} \times \Xi \to \mathbb{R}^n$ that given $w \in \mathcal{U}$ and $\xi \in \Xi$ returns an approximation $G(w,\xi) \in \mathbb{R}^n$ of a subgradient of f at w. The function G is such that

$$g(w) := \mathbb{E}_\Xi(G(w,\cdot)) \in \partial f(w) \quad \forall\, w \in \mathcal{U}.$$

d) There exists $M > 0$ such that G satisfies

$$\mathbb{E}_\Xi(\|G(w,\cdot)\|^2) \leq M^2 \quad \forall\, w \in \mathcal{U}.$$

Algorithm 4.2 (Stochastic subgradient method)**.**

0. Choose $w^1 \in \mathcal{W}$.

For $k = 1, 2, 3, \ldots$:

1. Draw ξ^k from $(\Xi, \mathbb{P}_\Xi)$ independently from $\xi^1, \ldots, \xi^{k-1}$ (this list is empty if $k = 1$) and compute $\hat{g}^k = G(w^k, \xi_k)$.

2. Choose a step size $\eta_k > 0$ and set $w^{k+1} = \pi_{\mathcal{W}}(w^k - \eta_k \hat{g}^k)$.

Example:

1. If f is defined by (25) then a standard choice is $(\Xi, \mathbb{P}_\Xi) = (\mathcal{Z}, \mathbb{P}_\mathcal{Z})$ and $G : \mathcal{U} \times \Xi \to \mathbb{R}^n$ is such that $G(w,\xi) \in \partial_w F(w,\xi)$.

   We then need a mechanism to draw independent samples from $(\mathcal{Z}, \mathbb{P}_\mathcal{Z})$.

2. If f is given by (26), we can choose $\Xi = \{z^{(1)}, \dots, z^{(N)}\}$ and $\mathbb{P}_\Xi$ such that

$$\mathbb{P}_\Xi(\xi) = \frac{|\{i \,:\, z^{(i)} = \xi\}|}{N} \quad \forall\, \xi \in \Xi.$$

   Drawing from $(\Xi, \mathbb{P}_\Xi)$ can be done by drawing uniformly an index j from $\{1, \dots, N\}$ and then returning $\xi = z^{(j)}$.

   A suitable function G is given by any function with $G(w, \xi) \in \partial_w F(w, \xi)$ for all $w \in \mathcal{U}$.

   A concrete realization of step 2 in the algorithm would be: Draw uniformly an index j_k from $\{1, \dots, N\}$, set $\xi^k = z^{(j_k)}$ and choose $\hat{g}^k \in \partial_w F(w^k, \xi^k) = \partial_w F(w^k, z^{(j_k)})$.

3. If f is again given by (26), we can choose $\Xi = \hat{\Xi}^q$, where $\hat{\Xi} = \{z^{(1)}, \dots, z^{(N)}\}$ and $\mathbb{P}_{\hat{\Xi}}$ are as Ξ and $\mathbb{P}_\Xi$ in 2. and $\mathbb{P}_\Xi = \mathbb{P}_{\hat{\Xi}} \times \cdots \times \mathbb{P}_{\hat{\Xi}}$.

   We then can choose

$$G(w, \xi) = \frac{1}{q} \sum_{i=1}^{q} \hat{G}(w, \xi_i),$$

   where $\xi = (\xi_1, \dots, \xi_q)$ and $\hat{G}$ is a function with $\hat{G}(w, \xi) \in \partial_w F(w, \xi)$ for all $w \in \mathcal{U}$.

   A concrete realization of step 2 in the algorithm would be:

   Draw uniformly and independently q indices $j_{k,1}, \dots, j_{k,q}$ from $\{1, \dots, N\}$, set $\xi^k = (z^{(j_{k,1})}, \dots, z^{(j_{k,q})})$, choose

$$\hat{g}^{k,i} \in \partial_w F(w^k, \xi_i^k) = \partial_w F(w^k, z^{(j_{k,q})})$$

   and set

$$\hat{g}^k = \frac{1}{q} \sum_{i=1}^{q} \hat{g}^{k,i}.$$

This version is called mini-batch stochastic gradient method.

We now prove the convergence of Algorithm 4.2, following [SZ13].

Theorem 4.3. *Let f be convex on the open convex neighborhood $\mathcal{U}$ of the closed convex set $\mathcal{W} \subset \mathbb{R}^n$. Let Assumption 4.1 hold and assume that there exists $\mu > 0$ such that the objective function is μ-strongly convex on $\mathcal{W}$, i.e.,*

$$f(\hat{w}) - f(w) \geq g^T(\hat{w} - w) + \frac{\mu}{2}\|\hat{w} - w\|^2 \quad \forall\, g \in \partial f(w), \quad \forall\, \hat{w}, w \in \mathcal{W}.$$

Denote by w^* the optimal solution of (24) *and let the sequence (w^k) be generated by Algorithm 4.2 with $\eta_k = 1/(\mu k)$. Then for all $k \geq 3$ there holds*

$$\mathbb{E}(f(w^k) - f(w^*)) \leq \frac{M^2(29 + 33\ln(K))}{2\mu k}.$$

Proof. We have for all $w \in \mathcal{W}$:

$$\begin{aligned} \|w^{k+1} - w\|^2 &= \|\pi_{\mathcal{W}}(w^k - \eta_k \hat{g}^k) - \pi_{\mathcal{W}}(w)\|^2 \leq \|w^k - \eta_k \hat{g}^k - w\|^2 \\ &= \|w^k - w\|^2 - 2\eta_k (\hat{g}^k)^T (w^k - w) + \eta_k^2 \|\hat{g}^k\|^2. \end{aligned} \tag{27}$$

We note that ξ^k is independent from $\xi^1, \ldots, \xi^{k-1}$ and that w^k only depends on $\xi^1, \ldots, \xi^{k-1}$. Hence, we have

$$\mathbb{E}(\|\hat{g}^k\|^2 | w^k) = \mathbb{E}(\|G(w^k, \xi^k)\|^2 | w^k) = \mathbb{E}_{\Xi}(\|G(w^k, \cdot)\|^2) \leq M^2,$$

which implies

$$\mathbb{E}(\|\hat{g}^k\|^2) \leq M^2.$$

In the same way, we obtain

$$\begin{aligned} \mathbb{E}((\hat{g}^k)^T (w^k - w)) &= \mathbb{E}(\mathbb{E}((\hat{g}^k)^T (w^k - w) | w^k)) = \mathbb{E}(\mathbb{E}_{\Xi}(G(w^k, \cdot)^T (w^k - w))) \\ &= \mathbb{E}((g^k)^T (w^k - w)), \end{aligned}$$

where $g^k = g(w^k) \in \partial f(w^k)$.

Thus, taking expectations in (27) yields

$$\begin{aligned} \mathbb{E}(\|w^{k+1} - w\|^2) &= \mathbb{E}(\|w^k - w\|^2) - 2\eta_k \mathbb{E}((\hat{g}^k)^T (w^k - w)) + \eta_k^2 \mathbb{E}(\|\hat{g}^k\|^2) \\ &\leq \mathbb{E}(\|w^k - w\|^2) - 2\eta_k \mathbb{E}((g^k)^T (w^k - w)) + \eta_k^2 M^2. \end{aligned} \tag{28}$$

For all $K, l \in \mathbb{N}$ with $l \leq k/2$ we get by summing over $k = K - l, \ldots, K$:

$$\begin{aligned} \mathbb{E}\Big(\sum_{k=K-l}^{K} (g^k)^T (w^k - w) \Big) &\leq \frac{1}{2\eta_{K-l}} \mathbb{E}(\|w^{K-l} - w\|^2) - \frac{1}{2\eta_K} \mathbb{E}(\|w^{K+1} - w\|^2) \\ &\quad + \frac{1}{2} \sum_{k=K-l+1}^{K} \left(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}} \right) \mathbb{E}(\|w^k - w\|^2) + \frac{M^2}{2} \sum_{k=K-l}^{K} \eta_k. \end{aligned}$$

Since $g^k \in \partial f(w^k)$, there holds

$$f(w) - f(w^k) \geq (g^k)^T (w - w^k).$$

Using this and $\eta_k = 1/(\mu k)$ yields:

$$\begin{aligned} \mathbb{E}\Big(\sum_{k=K-l}^{K} (f(w^k) - f(w)) \Big) &\leq \frac{\mu(K-l)}{2} \mathbb{E}(\|w^{K-l} - w\|^2) \\ &\quad + \frac{\mu}{2} \sum_{k=K-l+1}^{K} \mathbb{E}(\|w^k - w\|^2) + \frac{M^2}{2\mu} \sum_{k=K-l}^{K} \frac{1}{k}. \end{aligned} \tag{29}$$

We now are going to insert $w = w^{K-l}$. Before doing so, we use $\mathbb{E}(\|w^k - w^*\|^2) \leq \frac{4M^2}{\mu^2 k}$, see Lemma 4.4, and the fact that for all $u, v \in \mathbb{R}^n$, there holds

$$2\|u\|^2 + 2\|v\|^2 - \|u - v\|^2 = \|u\|^2 + 2\|v\|^2 + 2u^T v = \|u + v\|^2 \geq 0.$$

For $u = w^k - w^*$, $v = w^{K-l} - w^*$, $u - v = w^k - w^{K-l}$, $k \geq K - l$, this yields:

$$\begin{aligned}\mathbb{E}(\|w^k - w^{K-l}\|^2) &\leq 2\mathbb{E}(\|w^k - w^*\|^2) + 2\mathbb{E}(\|w^{K-l} - w^*\|^2) \\ &\leq \frac{8M^2}{\mu^2 k} + \frac{8M^2}{\mu^2(K-l)} \leq \frac{16M^2}{\mu^2(K-l)} \leq \frac{32M^2}{\mu^2 K}.\end{aligned}$$

Now, inserting $w = w^{K-l}$ into (29) and using this estimate gives

$$\mathbb{E}\Big(\sum_{k=K-l}^{K} (f(w^k) - f(w^{K-l}))\Big) \leq \frac{16M^2 l}{\mu K} + \frac{M^2}{2\mu}\sum_{k=K-l}^{K}\frac{1}{k}.$$

We use the abbreviation $S_l := \frac{1}{l+1}\sum_{k=K-l}^{K}\mathbb{E}(f(w^k))$. Then

$$-\mathbb{E}(f(w^{K-l})) \leq -\mathbb{E}(S_l) + \frac{M^2}{2\mu}\left(\frac{32}{K} + \sum_{k=K-l}^{K}\frac{1}{(l+1)k}\right)$$

Further,

$$\mathbb{E}(S_{l-1}) = \frac{l+1}{l}\mathbb{E}(S_l) - \frac{1}{l}\mathbb{E}(f(w^{K-l})) \leq \mathbb{E}(S_l) + \frac{M^2}{2\mu}\left(\frac{32}{lK} + \sum_{k=K-l}^{K}\frac{1}{l(l+1)k}\right)$$

We sum this over $l = 1, \ldots, \kappa$, where $\kappa = \lfloor K/2 \rfloor$. This gives

$$\mathbb{E}(f(w^K)) = \mathbb{E}(S_0) \leq \mathbb{E}(S_\kappa) + \frac{M^2}{2\mu}\sum_{l=1}^{\kappa}\left(\frac{32}{lK} + \sum_{k=K-l}^{K}\frac{1}{l(l+1)k}\right) \tag{30}$$

It remains to upper bound the right hand side. We will prove that for $K \geq 3$:

$$\mathbb{E}(S_\kappa) - f(w^*) \leq \tfrac{9M^2}{\mu K}, \tag{31}$$

$$\textstyle\sum_{l=1}^{\kappa}\frac{1}{l} \leq 1 + \ln(\kappa), \tag{32}$$

$$\textstyle\sum_{l=1}^{\kappa}\sum_{k=K-l}^{K}\frac{1}{l(l+1)k} \leq \frac{1}{K}(1 + \ln(K)). \tag{33}$$

Inserting these 3 estimates into (30) yields the assertion of the theorem:

$$\mathbb{E}(f(w^K)) - f(w^*) \leq \frac{9M^2}{\mu K} + \frac{16M^2}{\mu K}(1+\ln(\kappa)) + \frac{M^2}{2\mu K}(1+\ln(K)) \leq \frac{M^2(29 + 33\ln(K))}{2\mu K}.$$

To show (31), we set $w = w^*$ and $l = \kappa$ in (29) to obtain

$$\begin{aligned}\mathbb{E}(S_\kappa) - f(w^*) &\leq \frac{\mu(K-\kappa)}{2(\kappa+1)}\mathbb{E}(\|w^{K-\kappa} - w^*\|^2) \\ &\quad + \frac{\mu}{2(\kappa+1)}\sum_{k=K-\kappa+1}^{K}\mathbb{E}(\|w^k - w^*\|^2) + \frac{M^2}{2\mu(\kappa+1)}\sum_{k=K-\kappa}^{K}\frac{1}{k} \\ &\leq \frac{\mu(K-\kappa)}{2(\kappa+1)}\frac{4M^2}{\mu^2(K-\kappa)} + \frac{\mu}{2(\kappa+1)}\sum_{k=K-\kappa+1}^{K}\frac{4M^2}{\mu^2 k} + \frac{M^2}{2\mu(\kappa+1)}\sum_{k=K-\kappa}^{K}\frac{1}{k} \\ &\leq \frac{2M^2}{\mu(\kappa+1)} + \frac{2M^2}{\mu(\kappa+1)}\frac{\kappa}{K-\kappa+1} + \frac{M^2}{2\mu(\kappa+1)}\frac{\kappa+1}{K-\kappa} =: \frac{M^2}{2\mu}\cdot\frac{N}{D},\end{aligned}$$

where $D = (\kappa+1)(K-\kappa+1)(K-\kappa)$ and $N = 4(K-\kappa+1)(K-\kappa) + 4\kappa(K-\kappa) + (\kappa+1)(K-\kappa+1)$. For showing (31), we have to prove $N \leq \frac{18}{K}D$. Now if $K = 2\kappa$, then $K-\kappa = \kappa$ and thus $D = (\kappa+1)^2\kappa$. Further,

$$N = 4(\kappa+1)\kappa + 4\kappa^2 + (\kappa+1)^2 = 9\kappa^2 + 6\kappa + 1 \leq 9(\kappa+1)^2 \leq \frac{18}{K}D.$$

In the case $K = 2\kappa+1$ we get $D = (\kappa+1)^2(\kappa+2)$ and can estimate:

$$\begin{aligned} N &= 4(\kappa+2)(\kappa+1) + 4\kappa(\kappa+1) + (\kappa+1)(\kappa+2) = (\kappa+1)(9\kappa+10) \\ &\leq 9(\kappa+1)(\kappa+2) \leq \frac{9}{\kappa+1}D \leq \frac{18}{K}D. \end{aligned}$$

Therefore, (31) is proved.

Now, since $1/l \leq \int_{l-1}^{l} \frac{1}{t}\,dt$ for $l \geq 2$, there holds for all $2 \leq k_1 \leq k_2$

$$\sum_{l=k_1}^{k_2} \frac{1}{l} \leq \int_{k_1-1}^{k_2} \frac{1}{t}\,dt = \ln(k_2) - \ln(k_1-1).$$

Hence, $\sum_{l=1}^{\kappa} \frac{1}{l} \leq 1 + \ln(\kappa)$, which proves (32).

For showing (33), we set $\Sigma_K = \sum_{l=1}^{\kappa} \left(\frac{1}{l} + \frac{1}{K-l} \right)$ and note that

$$\sum_{l=1}^{\kappa} \sum_{k=K-l}^{K} \frac{1}{l(l+1)k} \leq \sum_{l=1}^{\kappa} \frac{1}{l(K-l)} = \frac{1}{K}\Sigma_K.$$

We show $\Sigma_K \leq 1 + \ln(K)$ for all $K \geq 3$:

$$\Sigma_3 = \frac{1}{1}+\frac{1}{2} = \frac{3}{2} < 2 < 1+\ln(3), \quad \Sigma_4 = \frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{2} = \frac{7}{3} \leq 2.34 < 2.38 < 1+\ln(4).$$

For $K \in \{2L, 2L+1\}$, $L \geq 2$, we have $\kappa = L$ and thus

$$\begin{aligned} \Sigma_{2L+1} - \Sigma_{2L} &= \frac{1}{2L} - \frac{1}{L} = -\frac{1}{2L} < 0, \\ \Sigma_{2L+2} - \Sigma_{2L} &= \frac{1}{L+1} + \frac{1}{2L+1} + \frac{1}{2L} - \frac{1}{L} \\ &= \frac{2L(2L+1) + 2L(L+1) - (L+1)(2L+1)}{2L(L+1)(2L+1)} \\ &= \frac{4L^2 + L - 1}{2L(L+1)(2L+1)} \leq \frac{1}{L+1} \leq \ln(L+1) - \ln(L). \end{aligned}$$

We can start an induction from $L = 2$ using $\Sigma_3, \Sigma_4 > 0$. For $2 \leq L \to L+1$ we obtain

$$\begin{aligned} &\Sigma_{2L+1} < \Sigma_{2L} \leq 1 + \ln(2L) < 1 + \ln(2L+1), \\ &\Sigma_{2L+2} \leq \Sigma_{2L} + \ln(L+1) - \ln(L) \leq 1 + \ln(2L) + \ln(L+1) - \ln(L) \\ &\qquad = 1 + \ln(2L(L+1)/L) = 1 + \ln(2L+2). \end{aligned}$$

This completes the induction and shows (33). □

Lemma 4.4. *Under the assumptions of Theorem 4.3, there holds*

$$\mathbb{E}(\|w^k - w^*\|^2) \leq \frac{4M^2}{\mu^2 k} \quad \forall\, k \in \mathbb{N}.$$

Proof. We use the strong convexity:

$$0 \geq f(w^*) - f(w^k) \geq (g^k)^T(w^* - w^k) + \frac{\mu}{2}\|w^k - w^*\|^2. \tag{34}$$

We show the assertion first for $k = 1$:

$$\|g^1\|\|w^1 - w^*\| \geq (g^1)^T(w^1 - w^*) \geq \frac{\mu}{2}\|w^1 - w^*\|^2.$$

Thus,

$$\|w^1 - w^*\|^2 \leq \frac{4}{\mu^2}\|g_1\|^2.$$

Further,

$$\begin{aligned}\mathbb{E}(\|\hat{g}^1\|^2) &= \mathbb{E}(\|g^1 + (\hat{g}^1 - g^1)\|^2) = \|g_1\|^2 + 2\mathbb{E}((g^1)^T(\hat{g}^1 - g^1)) + \mathbb{E}(\|\hat{g}^1 - g^1\|^2)\\ &\geq \|g_1\|^2 + 2\mathbb{E}((g^1)^T(\hat{g}^1 - g^1)) = \|g_1\|^2.\end{aligned}$$

Thus,

$$\|w^1 - w^*\|^2 \leq \frac{4}{\mu^2}\mathbb{E}(\|\hat{g}^1\|^2) \leq \frac{4M^2}{\mu^2}.$$

We now use that there exists $g^* \in \partial f(w^*)$ with

$$0 \leq f'(w^*, w^k - w^*) = (g^*)^T(w^k - w^*) \leq f(w^k) - f(w^*) - \frac{\mu}{2}\|w^k - w^*\|^2.$$

Using this, (28), and (34), we get

$$\begin{aligned}\mathbb{E}(\|w^{k+1} - w^*\|^2) &= \mathbb{E}(\|w^k - w^*\|^2) - 2\eta_k\mathbb{E}((g^k)^T(w^k - w^*)) + \eta_k^2 M^2\\ &\leq \mathbb{E}(\|w^k - w^*\|^2) - 2\eta_k\mathbb{E}(f(w^k) - f(w^*) + \tfrac{\mu}{2}\|w^k - w^*\|^2) + \eta_k^2 M^2\\ &\leq \mathbb{E}(\|w^k - w^*\|^2) - 2\eta_k\mathbb{E}(\tfrac{\mu}{2}\|w^k - w^*\|^2 + \tfrac{\mu}{2}\|w^k - w^*\|^2) + \eta_k^2 M^2\\ &\leq (1 - 2\eta_k\mu)\mathbb{E}(\|w^k - w^*\|^2) + \eta_k^2 M^2\\ &= (1 - 2\eta_k\mu)\mathbb{E}(\|w^k - w^*\|^2) + \eta_k^2 M^2\\ &= \big(1 - \tfrac{2}{k}\big)\mathbb{E}(\|w^k - w^*\|^2) + \frac{M^2}{\mu^2 k^2}.\end{aligned}$$

Inserting $k = 1$ shows that the assertion is true for $k = 2$:

$$\mathbb{E}(\|w^2 - w^*\|^2) \leq \frac{M^2}{\mu^2} \leq \frac{4M^2}{\mu^2 \cdot 2}.$$

We can now use induction: For $2 \leq k \to k + 1$ we obtain:

$$\mathbb{E}(\|w^{k+1} - w^*\|^2) \leq \big(1 - \tfrac{2}{k}\big)\frac{4M^2}{\mu^2 k} + \frac{M^2}{\mu^2 k^2} = \frac{4k - 7}{4k^2} \cdot \frac{4M^2}{\mu^2} \leq \frac{4M^2}{\mu^2(k + 1)},$$

since

$$\frac{1}{k+1} - \frac{4k-7}{4k^2} = \frac{4k^2 - (4k-7)(k+1)}{4k^2(k+1)} = \frac{3k+7}{4k^2(k+1)} > 0.$$

□

We note that on the way of proving convergence of function values of individual iterates, we also proved the following:

Setting $\overline{w}_l^K = \frac{1}{l+1}\sum_{K-l}^K w^k$, there holds by (31) and by the convexity of f:

$$\mathbb{E}(f(\overline{w}^K_{\lfloor K\rfloor})) - f(w^*) \leq \mathbb{E}(S_{\lfloor K\rfloor}) - f(w^*) \leq \frac{9M^2}{\mu K} = O(1/K).$$

Further, Lemma 4.4 shows

$$\mathbb{E}(\|w^K - w^*\|^2) \leq \frac{4M^2}{\mu^2 K} = O(1/K).$$

If we drop the assumption of μ-strict convexity, then an adaptation of the proof of Theorem 4.3 yields the following [SZ13, Thm. 2]:

Theorem 4.5. *Let f be convex on the open convex neighborhood $\mathcal{U}$ of the closed convex set $\mathcal{W} \subset \mathbb{R}^n$. Let Assumption 4.1 hold and assume that $D := \max_{w,w'\in\mathcal{W}} \|w - w'\| < \infty$. Denote by w^* the optimal solution of* (24) *and let the sequence (w^k) be generated by Algorithm 4.2 with $\eta_k = c/\sqrt{k}$, where $c > 0$ is fixed. Then for all $k \geq 2$ there holds*

$$\mathbb{E}(f(w^k)) - f(w^*) \leq \left(\frac{D^2}{c} + cM^2\right)\frac{2+\ln(k)}{\sqrt{k}}.$$

The estimate for individual iterates we derived are more recent and more difficult to obtain than estimates for the average $\overline{w}_{1,K}$. Improved versions for averaged iterates can be obtained if the average is taken only over the α-fraction $\alpha \in (0,1)$ of the most recent iterates. One can then show (assumming αK to be integer) in the setting of Theorem 4.3:

$$\mathbb{E}(f(w^K_{\alpha K})) - f(w^*) \leq \frac{17M^2(1 - \ln(\min\{\alpha, 1 + 1/T - \alpha\}))}{\mu K} = O(1/K).$$

In the setting of Theorem 4.5 one can show that the rate of $\mathbb{E}(f(w^K_{\alpha K})) - f(w^*)$ is at least $O(1/\sqrt{K})$.

The relatively slow rate of convergence is due to the nonsmoothness of the problem on the one hand and due to the "noise" in the (sub-) gradient approximations $\hat{g}^k$ on the other hand. Both of these challenges require step sizes η_k that converge to zero.

4.2 The case of convex functions with Lipschitz gradients

In the case where f has a Lipschitz gradient it is known that the deterministic gradient method with suitable constant step size achieves the rate

$$f(w^k) - f(w^*) = O(1/k).$$

There exist variants (Nesterov's accelerated gradient method, heavy ball method) where the rate is $O(1/k^2)$.

In the μ-strongly convex case and for a suitable constant step size the classical (deterministic) gradient method achieves a linear rate of convergence

$$f(w^k) - f(w^*) = O(\rho^k)$$

for some $\rho \in (0,1)$, which is faster than $O(1/k)$ or $O(1/k^2)$. Here, $\rho \approx 1 - \mu/L$ can be achieved.

In the case of accelerated gradient methods, one also achieves a linear rate of convergence, but ρ is smaller (which is better): $\rho \approx 1 - \sqrt{\mu/L}$.

We will now analyze the smooth situation in the more general stochastic setting.

In the following, we consider the unconstrained, strongly convex case with Lipschitz gradients:

$$\|\nabla f(w) - \nabla f(w')| \leq L\|w - w'\| \quad \forall\, w, w' \in \mathcal{M} \tag{35}$$

with a constant $L > 0$. Here, $\mathcal{M} \subset \mathcal{U}$ is a convex set that contains all iterates as well as the solution w^*.

From (35) we obtain for all $w, w + s \in \mathcal{M}$:

$$\begin{aligned} f(w+s) - f(w) &= \int_0^1 \nabla f(w+ts)^T s \, dt \\ &= \nabla f(w)^T s + \int_0^1 (\nabla f(w+ts) - \nabla f(w))^T s \, dt \\ &\leq \nabla f(w)^T s + \int_0^1 Lt \, dt \|s\|^2 = \nabla f(w)^T s + \frac{L}{2}\|s\|^2. \end{aligned} \tag{36}$$

At the solution w^* there holds $\nabla f(w^*) = 0$ and thus:

$$\|\nabla f(w)\| \leq \|\nabla f(w) - \nabla f(w^*)\| \leq L\|w - w^*\|.$$

Hence, using $w^k - w^{k+1} = \eta_k \nabla f(w^k)$, we see that step sizes $\eta_k \ll 1/L$ would result in $\|w^{k+1} - w^k\| \ll \|w^k - w^*\|$, which leads to slow convergence.

This shows that in the case of a μ-strongly convex f with Lipschitz gradients one should choose step sizes that are sufficiently small to ensure convergence, but bounded away from zero.

In a stochastic context, not the gradient is chosen, but an approximation. For achieving linear convergence, it is needed that the noise in this gradient estimate is attenuated sufficiently fast as k increases. Controlling this noise (the variance) is called *noise reduction* or *variance reduction*.

To formulate a suitable noise reduction variant of Assumption 4.1, d) we modify it appropriately. Also, for gaining some additional flexibility, we relax Assumption 4.1, c).

Assumption 4.6. As in Assumption 4.1 but with c), d) modified as follows:

With a convex set $\mathcal{M} \subset \mathcal{U}$ containing the solution and all iterates there holds:

c) Given are functions $G_k : \mathcal{U} \times \Xi \to \mathbb{R}^n$ such that there exist $0 < \alpha \leq \beta$ for which with $g_k(w) := \mathbb{E}_\Xi(G_k(w,\cdot))$ there holds for all $w \in \mathcal{M}$:

$$\|g_k(w)\| \leq \beta \|\nabla f(w)\|, \quad \nabla f(w)^T g_k(w) \geq \alpha \|\nabla f(w^k)\|^2.$$

d) There exist constants $M_1, M_2 \geq 0$, $0 < \zeta \leq 1$ such that

$$\mathbb{V}_\Xi(G_k(w,\cdot)) \leq M_1 \zeta^{k-1} + M_2 \|\nabla f(w)\|^2 \quad \forall\, w \in \mathcal{M},\ k \in \mathbb{N},$$

where

$$\begin{aligned}\mathbb{V}_\Xi(G_k(w,\cdot)) &:= \mathbb{E}_\Xi(\|G_k(w,\cdot)\|^2) - \|g_k(w)\|^2 \\ &= \mathbb{E}_\Xi(\|G_k(w,\cdot)\|^2) - \|\mathbb{E}_\Xi(G_k(w,\cdot))\|^2.\end{aligned}$$

Remark 4.7.

1. The requirements c) and d) are needed only at all $w = w^k$, $k \in \mathbb{N}$.
2. Instead of a joint probability space $(\Xi, \mathbb{P}_\Xi)$ we also could require that ξ^k is drawn from $(\Xi_k, \mathbb{P}_{\Xi_k})$ independently from $\xi^1, \ldots, \xi^{k-1}$. This, however can be unified by introducing a sufficiently big joint probability space Ξ. E.g., there could hold $\xi^k =\in \hat{\Xi}^{m_k}$ with $\mathbb{P}_{\Xi_k} = \mathbb{P}_{\hat{\Xi}} \times \cdots \times \mathbb{P}_{\hat{\Xi}}$ and a suitable choice for Ξ is then $\Xi = \hat{\Xi}^m$, where $m = \sup_k\ m_k$.
3. The conditions in c) allow, e.g., to choose $\hat{g}^k = B_k \tilde{g}^k$, where $\mathbb{E}(\tilde{g}^k | w^k) = \nabla f(w^k)$, ξ^k is independent from B_k and $B_k \in \mathbb{R}^{n\times n}$ is symmetric positive definite with all eigenvalues in $[\alpha, \beta]$. B_k can be used to approximate the inverse Hessian $\nabla^2 f(w^k)^{-1}$ in some sense.

We consider the following unconstrained stochastic gradient-type iteration:

$$w^{k+1} = w^k - \eta_k \hat{g}^k, \quad \hat{g}^k = G_k(w^k, \xi^k), \tag{37}$$

where ξ^k is drawn from $(\Xi, \mathbb{P}_\Xi)$ independently of ξ^j, $j < k$, and $\eta_k > 0$ is a deterministic step size.

We now derive all essential parts needed to establish a convergence result for (37). We work under Assumption 4.1, μ-strong convexity

$$f(\hat{w}) - f(w) \geq g^T(\hat{w} - w) + \frac{\mu}{2}\|\hat{w} - w\|^2 \quad \forall\, g \in \partial f(w), \quad \forall\, \hat{w}, w \in \mathcal{M}. \tag{38}$$

and the gradient Lipschitz condition (35).

With $\hat{g}^k = G_k(w^k, \xi^k)$ and $g^k = g(w^k)$ we can estimate:

$$\begin{aligned}\mathbb{E}(\|\hat{g}^k\|^2 | w^k) &= \mathbb{E}_\Xi(\|G_k(w^k,\cdot)\|^2) = \mathbb{V}_\Xi(G_k(w^k,\cdot)) + \|\mathbb{E}_\Xi(G_k(w^k,\cdot))\|^2 \\ &\leq M_1\zeta^{k-1} + M_2\|\nabla f(w^k)\|^2 + \|g^k\|^2 \\ &\leq M_1\zeta^{k-1} + (M_2 + \beta^2)\|\nabla f(w^k)\|^2 =: M_1\zeta^{k-1} + M_3\|\nabla f(w^k)\|^2.\end{aligned}$$

Now, using $w^{k+1} = w^k - \eta_k \hat{g}^k$:

$$f(w^{k+1}) - f(w^k) \leq \nabla f(w^k)^T (w^{k+1} - w^k) + \tfrac{L}{2}\|w^{k+1} - w^k\|^2 \leq -\eta_k \nabla f(w^k)^T \hat{g}^k + \tfrac{L}{2}\eta_k^2 \|\hat{g}^k\|^2,$$

and thus:

(39)
$$\begin{aligned}\mathbb{E}(f(w^{k+1})|w^k) - f(w^k) &\leq -\eta_k \nabla f(w^k)^T g^k + \tfrac{L}{2}\eta_k^2 (M_1 \zeta^{k-1} + M_3 \|\nabla f(w^k)\|^2) \\ &\leq -\alpha \eta_k \|\nabla f(w^k)\|^2 + \tfrac{L}{2}\eta_k^2 (M_1 \zeta^{k-1} + M_3 \|\nabla f(w^k)\|^2) \\ &\leq -\eta_k (\alpha - \tfrac{L}{2} M_3 \eta_k) \|\nabla f(w^k)\|^2 + \tfrac{LM_1}{2} \zeta^{k-1} \eta_k^2.\end{aligned}$$

For $\tau > 0$ and vectors b, c we obtain

$$0 \leq \|\tau b + (1/\tau) c\| = \tau^2 \|b^2\| + 2 b^T c + (1/\tau)^2 \|c\|^2.$$

Hence, with $\tau^2 = \mu$, $b = w^* - w$, and $c = \nabla f(w)$:

$$f(w^*) - f(w) \geq \nabla f(w)^T (w^* - w) + \frac{\mu}{2}\|w^* - w\|^2 \geq -\frac{1}{2\mu}\|\nabla f(w)\|^2.$$

Therefore,

$$-\|\nabla f(w^k)\|^2 \leq 2\mu (f(w^*) - f(w^k)).$$

This yields for $0 < \eta_k \leq \alpha/(LM_3)$:

$$\begin{aligned}\mathbb{E}(f(w^{k+1})|w^k) - f(w^k) &\leq -2\mu \eta_k (\alpha - \tfrac{L}{2} M_3 \eta_k)(f(w^k) - f(w^*)) + \tfrac{LM_1}{2} \zeta^{k-1} \eta_k^2 \\ &\leq -\mu \alpha \eta_k (f(w^k) - f(w^*)) + \tfrac{LM_1}{2} \zeta^{k-1} \eta_k^2.\end{aligned}$$

Hence,

$$\mathbb{E}(f(w^{k+1})|w^k) - f(w^*) \leq (1 - \alpha\mu\eta_k)(f(w^k) - f(w^*)) + \tfrac{LM_1}{2} \zeta^{k-1} \eta_k^2.$$

Taking expectation and choosing a constant step size $\eta_k = \eta \in (0, \alpha/(LM_3)]$ gives

$$\mathbb{E}(f(w^{k+1})) - f(w^*) \leq (1 - \alpha\mu\eta)(\mathbb{E}(f(w^k)) - f(w^*)) + \tfrac{LM_1}{2} \zeta^{k-1} \eta^2. \tag{40}$$

Based on this, we can derive the following convergence results:

Theorem 4.8. *Let f be convex on the open convex neighborhood $\mathcal{U}$ of the convex set $\mathcal{M} \subset \mathbb{R}^n$. Let Assumption 4.6 as well as* (35) *and* (38) *hold. Denote by w^* the unconstrained minimizer of f and let the sequence (w^k) be generated by the iteration* (37) *with constant step sizes $\eta_k = \eta \in (0, \alpha/(LM_3)]$. Assume that w^* and all w^k are contained in $\mathcal{M}$.*

Then in the case $\zeta < 1$, for any $\rho \in (1 - \alpha\mu\eta, 1)$ satisfying $\rho \geq \zeta$ and

$$C = \max\left(\tfrac{LM_1}{2(\rho + \alpha\mu\eta - 1)} \eta^2, f(w^1) - f(w^*)\right)$$

the following estimate holds:

$$\mathbb{E}(f(w^k)) - f(w^*) \leq C \rho^{k-1} \quad \forall\, k \in \mathbb{N}. \tag{41}$$

In the case $\zeta = 1$ (i.e., constant upper bound on the noise), there holds

$$\begin{aligned}\mathbb{E}(f(w^k)) - f(w^*) &\leq \frac{\eta L M_1}{2\alpha\mu} + (1 - \alpha\mu\eta)^{k-1} \left(f(w^1) - f(w^*) - \frac{\eta L M_1}{2\alpha\mu} \right) \\ &\to \frac{\eta L M_1}{2\alpha\mu} \leq \frac{M_1}{2\mu M_3} \quad (k \to \infty).\end{aligned} \tag{42}$$

Proof. Consider first the case $0 < \zeta < 1$.

With the given choice for C there holds $\mathbb{E}(f(w^1)) - f(w^*) = f(w^1) - f(w^*) \leq C = C\rho^0$.

For the induction step $k \to k+1$, assume $\mathbb{E}(f(w^k)) - f(w^*) \leq C\rho^{k-1}$. We observe that $1 - \alpha\mu\eta \geq 0$, which follows from $0 < \mu \leq L$ and $0 < \alpha \leq \beta$: $\alpha\mu\eta \leq \mu\alpha^2/(LM_3) \leq \alpha^2/M_3 \leq \alpha^2/\beta^2 \leq 1$.

Thus, from (40), we obtain:

$$\begin{aligned}\mathbb{E}(f(w^{k+1})) - f(w^*) &\leq (1-\alpha\mu\eta)C\rho^{k-1} + \tfrac{LM_1}{2}\eta^2\zeta^{k-1} \\ &\leq (1-\alpha\mu\eta)C\rho^{k-1} + C(\rho + \alpha\mu\eta - 1)\zeta^{k-1} \\ &\leq C[(1-\alpha\mu\eta) + (\rho + \alpha\mu\eta - 1)]\rho^{k-1} = C\rho^k.\end{aligned}$$

In the case $\zeta = 1$, (40) becomes

$$\mathbb{E}(f(w^{k+1})) - f(w^*) \leq (1-\alpha\mu\eta)(\mathbb{E}(f(w^k)) - f(w^*)) + \tfrac{LM_1}{2}\eta^2.$$

We note that

$$\tfrac{LM_1}{2}\eta^2 = \frac{\eta LM_1}{2\alpha\mu} - (1-\alpha\mu\eta)\frac{\eta LM_1}{2\alpha\mu}.$$

Hence

$$\mathbb{E}(f(w^{k+1})) - f(w^*) - \frac{\eta LM_1}{2\alpha\mu} \leq (1-\alpha\mu\eta)\left(\mathbb{E}(f(w^k)) - f(w^*) - \frac{\eta LM_1}{2\alpha\mu}\right),$$

which yields (42). □

For the deterministic gradient method, we can choose $M_1 = 0$, $M_2 = 0$, $\alpha = \beta = 1$ and obtain $0 < \eta \leq 1/L$ and

$$\rho = 1 - \eta\mu.$$

for $\eta = 1/L$ we get $\rho = 1 - \mu/L$ as stated above.

We have worked here with the Euclidean inner product, but it could be replaced by other inner products. Using the inner product $(v, w)_M = v^T M w$, where M is symmetric positive definite, results in a corresponding gradient representation $g_M(w) = M^{-1}\nabla f(w)$, since then $(g_M(w), s)_M = \nabla f(w)^T s$. The steepest descent direction w.r.t. $\|\cdot\|_M = (\cdot,\cdot)_M^{1/2}$ at w is then $-g_M(w) = -M^{-1}\nabla f(w)$. Thus, the gradient method with the M-inner product is the same as using $g^k = M^{-1}\nabla f(w^k)$ instead of $\nabla f(w^k)$ in the above method.

A well chosen matrix M (related to the Hessian $\nabla^2 f(w)$) can result in values of $\mu \leq L$ that are both close to one, where the requirements (35) and (38) are now

$$\begin{aligned}\|g_M(\hat{w}) - g_M(w)\|_M &\leq L\|\hat{w} - w\|_M, \\ f(\hat{w}) - f(w) &\geq (g_M(w), \hat{w} - w)_M + \frac{\mu}{2}\|\hat{w} - w\|_M^2\end{aligned}$$

for all $\hat{w}, w \in \mathcal{M}$.

Example: $f(w) = c^T w + \frac{1}{2} w^T Q w$, where Q is symmetric positive definite.

If we choose $M = Q$, then $g_M(w) = Q^{-1}(c + Qw) = Q^{-1}c + w$ and

$$\begin{aligned} f(\hat{w}) - f(w) - (g_M(w), \hat{w} - w)_M &= f(\hat{w}) - f(w) - \nabla f(w)^T(\hat{w} - w) \\ &= \frac{1}{2}(\hat{w} - w)^T Q(\hat{w} - w) = \frac{1}{2}\|\hat{w} - w\|_M^2. \end{aligned}$$

Thus, $\mu = 1$ can be chosen. Also,

$$\|g_M(\hat{w}) - g_M(w)\|_M = \|\hat{w} - w\|_M$$

and thus $L = 1$ can be chosen. The contraction factor in this case is $\rho = 0$, which means convergence in one step if no noise is present. Thus, if M is a good approximation to $\nabla^2 f(w) = Q$, then ρ is close to 0.

4.3 Complexity considerations

As a representative example, we consider (26), i.e.,

$$f(w) = \frac{1}{N} \sum_{i=1}^{N} F(w, z^{(i)}).$$

We measure computing time by the required number of evaluations of $\nabla F(w^k, z^{(i)})$.

The gradient method then requires time $O(N)$ per iteration. In the μ-strongly convex case with Lipschitz gradient, we can obtain the number of iterations to achieve accuracy ε by solving

$$C(1 - \mu/L)^k \leq \varepsilon$$

for k and obtain $k \geq \frac{\ln(C/\varepsilon)}{\ln(L/(L-\mu))} = O(\ln(1/\varepsilon))$. The required computing time is $O(N \ln(1/\varepsilon))$.

For the stochastic gradient method the computing time per iteration is τ and the required number of iterations is obtained from

$$\frac{C}{k} \leq \varepsilon.$$

Thus $O(1/\varepsilon)$ iterations are needed and the computing time is $O(1/\varepsilon)$.

If, e.g., $N = 10^6$ and $\varepsilon = 10^{-3}$, then

$$\frac{1}{\varepsilon} = 10^3, \quad N \ln(1/\varepsilon) = 10^6 \cdot 3 \cdot \ln(10) \approx 7 \cdot 10^6.$$

Thus, modulo constant factors, the stochastic gradient method is preferable.

If, e.g., $N = 10^3$, $\varepsilon = 10^{-5}$, then

$$\frac{1}{\varepsilon} = 10^5, \quad N \ln(1/\varepsilon) = 10^3 \cdot 5 \cdot \ln(10) \approx 11.5 \cdot 10^3.$$

Here, the gradient method can be expected to be faster.

4.4 Noise reduction

To introduce the idea of noise reduction techniques, we mainly focus on mini-batch stochastic gradient methods with dynamic mini-batch sizes.

Dynamic adjustment of mini-batch size

A possibility to control the variance is by dynamically adjusting the mini-batch size m_k and to use

$$G_k(w,\xi) = \frac{1}{m_k}\sum_{j=1}^{m_k} \nabla_w F(w,\xi^j).$$

Here, $\xi^j = z^{(i_j)}$, $1 \le j \le m_k$, and $i^1,\dots,i^{m_k}$ are drawn from $\{1,\dots,N\}$ uniformly and independently.

There then holds

$$\mathbb{E}(\nabla_w F(w,\xi^j)) = \frac{1}{N}\sum_{i=1}^{N} \nabla_w F(w,z^{(i)}) = \nabla f(w),$$

$$\mathbb{E}(\|\nabla_w F(w,\xi^j)\|^2) = \frac{1}{N}\sum_{i=1}^{N} \|\nabla_w F(w,z^{(i)})\|^2 =: \overline{E^2},$$

and thus:

$$\begin{aligned}
\mathbb{E}(\|G_k(w,\cdot)\|^2) &= \frac{1}{m_k^2}\sum_{j=1}^{m_k}\sum_{l=1}^{m_k} \mathbb{E}(\nabla_w F(w,\xi^j)^T \nabla_w F(w,\xi^l)) \\
&= \frac{1}{m_k^2}\sum_{j=1}^{m_k} \mathbb{E}(\|\nabla_w F(w,\xi^j)\|^2 + \frac{2}{m_k^2}\sum_{j=1}^{m_k-1}\sum_{l=j+1}^{m_k} \mathbb{E}(\nabla_w F(w,\xi^j)^T \nabla_w F(w,\xi^l)) \\
&= \frac{1}{m_k}\overline{E^2} + \frac{m_k(m_k-1)}{m_k^2}\nabla f(w)^T\nabla f(w) = \frac{1}{m_k}\overline{E^2} + (1-\frac{1}{m_k^2})\|\nabla f(w)\|^2.
\end{aligned}$$

This yields

$$\mathbb{V}(G_k(w,\cdot)) = \mathbb{E}(\|G_k(w,\cdot)\|^2) - \|\nabla f(w)\|^2 = \frac{1}{m_k}(\overline{E^2} - \|\nabla f(w)\|^2),$$

showing that the variance $\mathbb{V}(G_k(w,\cdot))$ is proportional to $1/m_k$.

Let us consider the iteration (37) with $\alpha = \beta = 1$, $M_2 = 0$, $M_1 > 0$, $\eta = 1/L$, $\zeta = \rho = 1 - \mu/L$.

To achieve the bound on the variance, we have to increase the mini-batch size according to $m_k = O(1/\zeta^{k-1}) = O(1/\rho^{k-1})$. Then the work in τ-units for obtaining w^{k+1} is

$$\sum_{l=1}^{k} m_k = \sum_{l=1}^{k} \rho^{1-l} = \frac{1-\rho^{-k}}{1-\rho^{-1}} = \frac{\rho^{1-k}-\rho}{1-\rho} \le \frac{\rho^{1-k}}{1-\rho} = \frac{L}{\mu}\rho^{1-k} = O(\rho^{1-k}).$$

To achieve $\rho^k \le \varepsilon$, the required work is approximately $O(1/\varepsilon)$,where we have used $\rho^{1-k} \le \rho^{-k}$.

Improvements can be achieved by not choosing m_k according to a rule that uses an a-priori bound for the variance, but to estimate the variance, e.g., using sample averages, and adjust m_k accordingly.

Stochastic variance reduced gradient (SVRG)

Other noise reduction methods use the idea to draw i_k randomly from $\{1, \ldots, N\}$ and to make, instead of $\nabla_w F(w^k, z^{(i_k)})$, the choice

$$\hat{g}^k = \nabla_w F(w^k, z^{(i_k)}) - \nabla_w F(\tilde{w}^s, z^{(i_k)}) + \tilde{g}^s. \tag{43}$$

Here, $\tilde{w}^s$ is an approximation to w^* that is computed from previous iterates and

$$\tilde{g}^s = g(\tilde{w}^s) = \frac{1}{N} \sum_{i=1}^{N} \nabla_w F(\tilde{w}^s, z^{(i)}).$$

We note that the modification term has expected value 0, since

$$\mathbb{E}(\nabla_w F(\tilde{w}^s, z^{(i_k)}) | \tilde{w}^s) = \tilde{g}^s.$$

The method uses an outer iteration $s = 1, 2, \ldots$. For outer iteration s, $\tilde{w}^s$ and $\tilde{g}^s$ are computed and then m inner iterations $k = (s-1)m + 1, \ldots, sm$ are performed for fixes s.

Options for choosing $\tilde{w}^{s+1}$ are:

- $\tilde{w}^{s+1} = w^{sm} =$ final inner iterate of outer iteration s

  or

- $\tilde{w}^{s+1} = w^{(s-1)m+l_s}$, where l_s is randomly chosen from $\{1, \ldots, m\}$.

The resulting method is called SVRG (stochastic variance reduced gradient). One can show linear convergence on expectation if m is sufficiently large.

The underlying idea is that the choice (43) of $\hat{g}^k$ has a smaller variance than the usual choice $\nabla_w F(w^k, z^{(i_k)})$. This can be, e.g., seen from the fact that if $\tilde{w}^s \to w^*$ and $w^k \to w^*$, then

$$\begin{aligned} &\tilde{g}^s = g(\tilde{w}^s) \to g(w^*) = 0, \\ &\nabla_w F(w^k, z^{(i)}) - \nabla_w F(\tilde{w}^s, z^{(i)}) \to \nabla_w F(w^*, z^{(i)}) - \nabla_w F(w^*, z^{(i)}) = 0 \end{aligned}$$

for all i if $\nabla_w F$ is contiuous w.r.t. w.

On the other hand,

$$\nabla_w F(w^k, z^{(i)}) \to \nabla_w F(w^*, z^{(i)})$$

and although $\nabla f(w^*) = 0$, there usually holds $\nabla_w F(w^*, z^{(i)}) \neq 0$ for most or all i.

4.5 The non-convex case

Based on our previous results, we can also develop convergence results for the non-convex case quite immediately. This is important, e.g., in deep learning, since neural networks are nonlinear and result in non-convex risk minimization problems.

We assume that the cost function $f : \mathcal{U} \to \mathbb{R}$ has a Lipschitz continuous gradient but f is allowed to be *non-convex*.

We can derive (39) as before and taking expectation yields

$$\mathbb{E}(f(w^{k+1})) - \mathbb{E}(f(w^k)) \leq -\eta_k(\alpha - \tfrac{L}{2}M_3\eta_k)\mathbb{E}(\|\nabla f(w^k)\|^2) + \tfrac{LM_1}{2}\zeta^{k-1}\eta_k^2.$$

For $\eta_k \leq \alpha/(LM_3)$ we get

$$\mathbb{E}(f(w^{k+1})) - \mathbb{E}(f(w^k)) \leq -\tfrac{1}{2}\alpha\mathbb{E}(\|\nabla f(w^k)\|^2) + \tfrac{LM_1}{2}\zeta^{k-1}\eta_k^2.$$

Assuming that $f^* := \inf_{x\in\mathcal{U}} f(x)$ is finite, we obtain by summing over $k = 1, \ldots, K$:

$$f^* - f(w^1) \leq \mathbb{E}(f(w^{K+1})) - f(w^1) \leq -\frac{\alpha}{2}\sum_{k=1}^{K}\eta_k\mathbb{E}(\|\nabla f(w^k)\|^2) + \tfrac{LM_1}{2}\sum_{k=1}^{K}\zeta^{k-1}\eta_k^2.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left(\sum_{k=1}^{K}\eta_k\|\nabla f(w^k)\|^2\right) &= \sum_{k=1}^{K}\eta_k\mathbb{E}(\|\nabla f(w^k)\|^2) \\ &\leq \frac{2(f(w^1) - f^*)}{\alpha} + \tfrac{LM_1}{\alpha}\sum_{k=1}^{K}\zeta^{k-1}\eta_k^2.\end{aligned} \tag{44}$$

From this, we can obtain several flavors of convergence results from this:

- If $\sum_{k=1}^{\infty}\zeta^{k-1}\eta_k^2 < \infty$, then there exists $C > 0$ with
$$\mathbb{E}\left(\sum_{k=1}^{\infty}\eta_k\|\nabla f(w^k)\|^2)\right) \leq C.$$
- Further, if in addition $\sum_{k=1}^{\infty}\eta_k = \infty$, then there holds
$$\begin{aligned}&\liminf_{k\to\infty}\mathbb{E}(\|\nabla f(w^k)\|^2) = 0, \\ &\mathbb{E}\left(\frac{1}{\sum_{k=1}^{K}\eta_k}\sum_{k=1}^{K}\eta_k\|\nabla f(w^k)\|^2)\right) \to 0 \quad (K \to \infty).\end{aligned}$$
The lim inf-result follows from the fact that otherwise, there would exist $\delta > 0$ and $k' \geq 1$ with $\mathbb{E}(\|\nabla f(w^k)\|^2) \geq \delta$ for all $k \geq k'$, which results in the contradiction
$$C \geq \mathbb{E}\left(\sum_{k=1}^{\infty}\eta_k\|\nabla f(w^k)\|^2)\right) \geq \delta\sum_{k=k'}^{\infty}\eta_k = \infty.$$
The second result follows easily from a) since $\sum_{k=1}^{K}\eta_k \to 0$.

If $\eta_k = \eta \in (0, \alpha/(LM_3)]$ is fixed, then (44) becomes:

$$\begin{aligned}\mathbb{E}\left(\sum_{k=1}^{K} \|\nabla f(w^k)\|^2\right) &\leq \frac{2(f(w^1) - f^*)}{\alpha\eta} + \tfrac{LM_1\eta}{\alpha} \sum_{k=1}^{K} \zeta^{k-1} \\ &\leq \frac{2(f(w^1) - f^*)}{\alpha\eta} + \begin{cases} \frac{LM_1\eta}{\alpha(1-\zeta)} & (0 < \zeta < 1) \\ \frac{LM_1\eta K}{\alpha} & (\zeta = 1). \end{cases}\end{aligned}$$

- In the case $0 < \zeta < 1$ we can conclude that there exists $C > 0$ with

  $$\mathbb{E}\left(\sum_{k=1}^{\infty} \|\nabla f(w^k)\|^2\right) \leq C,$$

  which implies

  $$\mathbb{E}\left(\frac{1}{K}\sum_{k=1}^{K} \|\nabla f(w^k)\|^2\right) \leq \frac{C}{K},$$

  and also $\mathbb{E}(\|\nabla f(w^k)\|^2) \to 0$, $k \to \infty$.

- In the case $\zeta = 1$ there holds

  $$\mathbb{E}\left(\frac{1}{K}\sum_{k=1}^{K} \|\nabla f(w^k)\|^2\right) \leq \frac{2(f(w^1) - f^*)}{\alpha\eta K} + \tfrac{LM_1\eta}{\alpha} \to \tfrac{LM_1\eta}{\alpha} \quad (K \to \infty).$$

Except for the very last case, we see that we obtain results that can be interpreted as convergence to stationarity in expectation.

References

[BCN16] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, Technical Report 1606.04838, arXiv, 2016.

[GI06] T. Glasmachers, C. Igel, *Maximum-gain working set selection for SVMs*, Journal of Machine Learning Research 7, 1437–1466, 2006.

[Hoe63] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association 58, 13–30, 1963.

[LS04] N. List and H. U. Simon, *A general convergence theorem for the decomposition method*. In: J. Shawe-Taylor and Y. Singer, eds., Proceedings of the 17th Annual Conference on Learning Theory, COLT 2004, LNCS 3120, 363–377, Springer-Verlag, 2004.

[Vap95] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, 1995.

[Vap99] V. N. Vapnik, *An overview of statistical learning theory*, IEEE Transactions on Neural Networks 10, 988–999, 1999.