INTRODUCTION TO MACHINE LEARNING

Introduction to Machine Learning

Alex Smola and S.V.N. Vishwanathan

Yahoo! Labs
Santa Clara
–and–
Departments of Statistics and Computer Science
Purdue University
–and–
College of Engineering and Computer Science
Australian National University

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011–4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2008

Printed in the United Kingdom at the University Press, Cambridge

Typeface Monotype Times 10/13pt *System* LaTeX 2ε [Alexander J. Smola and S.V.N. Vishwanathan]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data available

ISBN 0 521 82583 0 hardback

Author:
Revision:
Timestamp: June 10, 2015
URL:

Contents

1

Online Learning and Boosting

So far the learning algorithms we considered assumed that all the training data is available before building a model for predicting labels on unseen data points. In many modern applications data is available only in a streaming fashion, and one needs to predict labels on the fly. To describe a concrete example, consider the task of spam filtering. As emails arrive the learning algorithm needs to classify them as spam or ham. Tasks such as these are tackled via online learning. Online learning proceeds in rounds. At each round a training example is revealed to the learning algorithm, which uses its current model to predict the label. The true label is then revealed to the learner which incurs a loss and updates its model based on the feedback provided. This protocol is summarized in Algorithm 1.1. The goal of online learning is to minimize the total loss incurred. By an appropriate choice of labels and loss functions, this setting encompasses a large number of tasks such as classification, regression, and density estimation. In our spam detection example, if an email is misclassified the user can provide feedback which is used to update the spam filter, and the goal is to minimize the number of misclassified emails.

1.1 Halving Algorithm

The halving algorithm is conceptually simple, yet it illustrates many of the concepts in online learning. Suppose we have access to a set of n experts, that is, functions f_i which map from the input space $\mathcal{X}$ to the output space $\mathcal{Y} = \{\pm 1\}$. Furthermore, assume that one of the experts is consistent, that is, there exists a $j \in \{1, \ldots, n\}$ such that $f_j(x_t) = y_t$ for $t = 1, \ldots, T$. The halving algorithm maintains a set $\mathcal{C}_t$ of consistent experts at time t. Initially $\mathcal{C}_0 = \{1, \ldots, n\}$, and it is updated recursively as

$$\mathcal{C}_{t+1} = \{i \in \mathcal{C}_t \text{ s.t. } f_i(x_{t+1}) = y_{t+1}\}. \tag{1.1}$$

The prediction on a new data point is computed via a majority vote amongst the consistent experts: $\hat{y}_t = \text{majority}(\mathcal{C}_t)$.

Lemma 1.1 *The Halving algorithm makes at most* $\log_2(n)$ *mistakes.*

Algorithm 1.1 Protocol of Online Learning

```
for $t = 1, \ldots, T$ do do
    Get training instance $x_t$
    Predict label $\hat{y}_t$
    Get true label $y_t$
    Incur loss $l(\hat{y}_t, x_t, y_t)$
    Update model
end for
```

Proof Let M denote the total number of mistakes. The halving algorithm makes a mistake at iteration t if at least half the consistent experts $\mathcal{C}_t$ predict the wrong label. This in turn implies that

$$|\mathcal{C}_{t+1}| \leq \frac{|\mathcal{C}_t|}{2} \leq \frac{|\mathcal{C}_0|}{2^M} = \frac{n}{2^M}.$$

On the other hand, since one of the experts is consistent it follows that $1 \leq |\mathcal{C}_{t+1}|$. Therefore, $2^M \leq n$. Solving for M completes the proof. ■

1.2 Weighted Majority

We now turn to the scenario where none of the experts is consistent. Therefore, the aim here is not to minimize the number mistakes but to minimize regret.

1.3 Stochastic Mirror Descent

In this section we will consider optimization algorithms for solving the following problem:

$$\min_{w \in \Omega} J(w) \text{ where } J(w) = \sum_{t=1}^{T} f_t(w). \tag{1.2}$$

Suppose we have access to a function ψ which is continuously differentiable and strongly convex with modulus of strong convexity $\sigma > 0$ (see Section **??** for definition of strong convexity). The diameter of Ω as measured by the Bregman divergence Δ_ψ (see Section **??** for the definition and important properties of Bregman divergences) is given by

$$\operatorname{diam}_\psi(\Omega) = \max_{w, w' \in \Omega} \Delta_\psi(w, w'). \tag{1.3}$$

Algorithm 1.2 Stochastic Mirror Descent

```
**Input:** Initial point $w_1$, maximum iterations $T$
**for** $t = 1, \ldots, T$ **do**
    Compute $\hat{w}_{t+1} = \nabla\psi^* \left(\nabla\psi(w_t) - \eta_t g_t\right)$ with $g_t := \partial_w f_t(w_t)$
    Set $w_{t+1} = P_{\psi,\Omega}\left(\hat{w}_{t+1}\right)$
**end for**
**Return:** $w_{T+1}$
```

For the rest of this section we will make the following standard assumptions:

- Each f_t is convex and revealed at time instance t.
- Ω is a closed convex subset of $\mathbb{R}^n$ with non-empty interior.
- The diameter $\operatorname{diam}_\psi(\Omega)$ of Ω is bounded by $F < \infty$.
- The set of optimal solutions of (1.2) denoted by Ω^* is non-empty.
- The subgradient $\partial_w f_t(w)$ can be computed for every t and $w \in \Omega$.
- The Bregman projection (**??**) can be computed for every $w' \in \mathbb{R}^n$.
- The gradient $\nabla\psi$, and its inverse $(\nabla\psi)^{-1} = \nabla\psi^*$ can be computed.

The method we employ to solve (1.2) is given in Algorithm 1.2.

Our key result is Lemma 1.2 given below. It can be found in various guises in different places most notably Lemma 2.1 and 2.2 in [Ned02], Theorem 4.1 and Eq. (4.21) and (4.15) in [BT03], in the proof of Theorem 1 of [Zin03], as well as Lemma 3 of [SSS07]. We prove a slightly general variant; we allow for projections with an arbitrary Bregman divergence and also take into account the generalized version of strong convexity of f_t. Both these modifications will allow us to deal with general settings within a unified framework.

Lemma 1.2 *Let f_t be strongly convex with respect to ψ with modulus $\lambda \geq 0$ for all t. For any $w \in \Omega$ the sequences generated by Algorithm 1.2 satisfy*

$$\Delta_\psi(w, w_{t+1}) \leq \Delta_\psi(w, w_t) - \eta_t \left\langle g_t, w_t - w \right\rangle + \frac{\eta_t^2}{2\sigma} \left\| g_t \right\|^2 \tag{1.4}$$

$$\leq (1 - \eta_t \lambda)\Delta_\psi(w, w_t) - \eta_t(f_t(w_t) - f_t(w)) + \frac{\eta_t^2}{2\sigma} \left\| g_t \right\|^2 . \tag{1.5}$$

Proof We prove the result in three steps. First we upper bound $\Delta_\psi(w, w_{t+1})$ by $\Delta_\psi(w, \hat{w}_{t+1})$. This is a consequence of (**??**) and the non-negativity of the Bregman divergence which allows us to write

$$\Delta_\psi(w, w_{t+1}) \leq \Delta_\psi(w, \hat{w}_{t+1}). \tag{1.6}$$

In the next step we use Lemma **??** to write

$$\Delta_\psi(w, w_t) + \Delta_\psi(w_t, \hat{w}_{t+1}) - \Delta_\psi(w, \hat{w}_{t+1}) = \left\langle \nabla\psi(\hat{w}_{t+1}) - \nabla\psi(w_t), w - w_t \right\rangle .$$

Since $\nabla\psi^* = (\nabla\psi)^{-1}$, the update in step 3 of Algorithm 1.2 can equivalently be written as $\nabla\psi(\hat{w}_{t+1}) - \nabla\psi(w_t) = -\eta_t g_t$. Plugging this in the above equation and rearranging

$$\Delta_\psi(w, \hat{w}_{t+1}) = \Delta_\psi(w, w_t) - \eta_t \langle g_t, w_t - w\rangle + \Delta_\psi(w_t, \hat{w}_{t+1}). \tag{1.7}$$

Finally we upper bound $\Delta_\psi(w_t, \hat{w}_{t+1})$. For this we need two observations: First, $\langle x, y\rangle \leq \frac{1}{2\sigma}\|x\|^2 + \frac{\sigma}{2}\|y\|^2$ for all $x, y \in \mathbb{R}^n$ and $\sigma > 0$. Second, the σ strong convexity of ψ allows us to bound $\Delta_\psi(\hat{w}_{t+1}, w_t) \geq \frac{\sigma}{2}\|w_t - \hat{w}_{t+1}\|^2$. Using these two observations

$$\begin{aligned}
\Delta_\psi(w_t, \hat{w}_{t+1}) &= \psi(w_t) - \psi(\hat{w}_{t+1}) - \langle\nabla\psi(\hat{w}_{t+1}), w_t - \hat{w}_{t+1}\rangle \\
&= -(\psi(\hat{w}_{t+1}) - \psi(w_t) - \langle\nabla\psi(w_t), \hat{w}_{t+1} - w_t\rangle) + \langle\eta_t g_t, w_t - \hat{w}_{t+1}\rangle \\
&= -\Delta_\psi(\hat{w}_{t+1}, w_t) + \langle\eta_t g_t, w_t - \hat{w}_{t+1}\rangle \\
&\leq -\frac{\sigma}{2}\|w_t - \hat{w}_{t+1}\|^2 + \frac{\eta_t^2}{2\sigma}\|g_t\|^2 + \frac{\sigma}{2}\|w_t - \hat{w}_{t+1}\|^2 \\
&= \frac{\eta_t^2}{2\sigma}\|g_t\|^2. \qquad (1.8)
\end{aligned}$$

Inequality (1.4) follows by putting together (1.6), (1.7), and (1.8), while (1.5) follows by using (**??**) with $f = f_t$ and $w' = w_t$ and substituting into (1.4). ■

Now we are ready to prove regret bounds.

Lemma 1.3 *Let $w^* \in \Omega^*$ denote the best parameter chosen in hindsight, and let $\|g_t\| \leq L$ for all t. Then the regret of Algorithm 1.2 can be bounded via*

$$\sum_{t=1}^T f_t(w_t) - f_t(w^*) \leq F\left(\frac{1}{\eta_T} - T\lambda\right) + \frac{L^2}{2\sigma}\sum_{t=1}^T \eta_t. \tag{1.9}$$

Proof Set $w = w^*$ and rearrange (1.5) to obtain

$$f_t(w_t) - f_t(w^*) \leq \frac{1}{\eta_t}\left((1 - \lambda\eta_t)\Delta_\psi(w^*, w_t) - \Delta_\psi(w^*, w_{t+1})\right) + \frac{\eta_t}{2\sigma}\|g_t\|^2.$$

Summing over t

$$\sum_{t=1}^T f_t(w_t) - f_t(w^*) \leq \underbrace{\sum_{t=1}^T \frac{1}{\eta_t}\left((1 - \eta_t\lambda)\Delta_\psi(w^*, w_t) - \Delta_\psi(w^*, w_{t+1})\right)}_{T_1} + \underbrace{\sum_{t=1}^T \frac{\eta_t}{2\sigma}\|g_t\|^2}_{T_2}.$$

Since the diameter of Ω is bounded by F and Δ_ψ is non-negative

$$\begin{aligned}
T_1 &= \left(\frac{1}{\eta_1} - \lambda\right) \Delta_\psi(w^*, w_1) - \frac{1}{\eta_T}\Delta_\psi(w^*, w_{T+1}) + \sum_{t=2}^{T} \Delta_\psi(w^*, w_t) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \lambda\right) \\
&\leq \left(\frac{1}{\eta_1} - \lambda\right) \Delta_\psi(w^*, w_1) + \sum_{t=2}^{T} \Delta_\psi(w^*, w_t) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \lambda\right) \\
&\leq \left(\frac{1}{\eta_1} - \lambda\right) F + \sum_{t=2}^{T} F \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \lambda\right) = F \left(\frac{1}{\eta_T} - T\lambda\right).
\end{aligned}$$

On the other hand, since the subgradients are Lipschitz continuous with constant L it follows that

$$T_2 \leq \frac{L^2}{2\sigma} \sum_{t=1}^{T} \eta_t.$$

Putting together the bounds for T_1 and T_2 yields (1.9). ■

Corollary 1.4 *If $\lambda > 0$ and we set $\eta_t = \frac{1}{\lambda t}$ then*

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \frac{L^2}{2\sigma\lambda}(1 + \log(T)),$$

On the other hand, when $\lambda = 0$, if we set $\eta_t = \frac{1}{\sqrt{t}}$ then

$$\sum_{t=1}^{T} f_t(x_t) - f_t(x^*) \leq \left(F + \frac{L^2}{\sigma}\right)\sqrt{T}.$$

Proof First consider $\lambda > 0$ with $\eta_t = \frac{1}{\lambda t}$. In this case $\frac{1}{\eta_T} = T\lambda$, and consequently (1.9) specializes to

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w^*) \leq \frac{L^2}{2\sigma\lambda} \sum_{t=1}^{T} \frac{1}{t} \leq \frac{L^2}{2\sigma\lambda}(1 + \log(T)).$$

When $\lambda = 0$, and we set $\eta_t = \frac{1}{\sqrt{t}}$ and use problem 1.2 to rewrite (1.9) as

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w^*) \leq F\sqrt{T} + \frac{L^2}{\sigma} \sum_{t=1}^{T} \frac{1}{2\sqrt{t}} \leq F\sqrt{T} + \frac{L^2}{\sigma}\sqrt{T}.$$

■

1.3.1 Adaptive Learning Rates

Lemma 1.5 *For any $w \in \Omega$ the sequences generated by Algorithm 1.2 using*

- $\eta_t = \eta$ *for all* t*, and*
- $\psi_t = \frac{1}{2} \left\|\cdot\right\|_{H_t}^2$ *instead of* ψ*, where* $H_t := diag(s_t)$ *and* $s_{t,i} := \sqrt{\sum_{j=1}^t g_{j,i}^2}$

satisfy

$$\Delta_{\psi_t}(w, w_{t+1}) \le \Delta_{\psi_t}(w, w_t) - \eta(f_t(w_t) - f_t(w)) + \frac{\eta^2}{2} \left\|g_t\right\|_{H_t^*}^2 \tag{1.10}$$

where $H_t^* = H_t^{-1}$.

Proof The proof is the same as that for Lemma 1.2, except for the step where we upper bound $\Delta_\psi(w_t, \hat{w}_{t+1})$. We use the Fenchel Young inequality to write $\langle x, y\rangle \le \frac{1}{2} \left\|x\right\|_{H_t}^2 + \frac{1}{2} \left\|y\right\|_{H_t^*}^2$ for all $x, y \in \mathbb{R}^n$. Moreover, $\Delta_{\psi_t}(\hat{w}_{t+1}, w_t) = \frac{1}{2} \left\|w_t - \hat{w}_{t+1}\right\|_{H_t}^2$. Using these two observations

$$\begin{aligned}
\Delta_\psi(w_t, \hat{w}_{t+1}) &= \psi(w_t) - \psi(\hat{w}_{t+1}) - \langle \nabla\psi(\hat{w}_{t+1}), w_t - \hat{w}_{t+1}\rangle \\
&= -(\psi(\hat{w}_{t+1}) - \psi(w_t) - \langle \nabla\psi(w_t), \hat{w}_{t+1} - w_t\rangle) + \langle \eta g_t, w_t - \hat{w}_{t+1}\rangle \\
&= -\Delta_\psi(\hat{w}_{t+1}, w_t) + \langle \eta g_t, w_t - \hat{w}_{t+1}\rangle \\
&\le -\Delta_\psi(\hat{w}_{t+1}, w_t) + \frac{\eta^2}{2} \left\|g_t\right\|_{H_t^*}^2 + \frac{1}{2} \left\|w_t - \hat{w}_{t+1}\right\|_{H_t}^2 \\
&= \frac{\eta^2}{2} \left\|g_t\right\|_{H_t^*}^2 .
\end{aligned}$$

■

As before, let $w^* \in \Omega^*$ denote the best parameter chosen in hindsight. Letting $w = w^*$ in (1.10), rearranging, and summing over t obtains

$$\sum_{t=1}^T f_t(w_t) - f_t(w^*) \le \underbrace{\sum_{t=1}^T \frac{1}{\eta} (\Delta_{\psi_t}(w^*, w_t) - \Delta_{\psi_t}(w^*, w_{t+1}))}_{T_1} + \underbrace{\sum_{t=1}^T \frac{\eta}{2} \left\|g_t\right\|_{H_t^*}^2}_{T_2} .$$

Since Δ_{ψ_t} is non-negative, we have

$$\begin{aligned}
T_1 &= \frac{1}{\eta}\Delta_{\psi_1}(w^*, w_1) - \frac{1}{\eta}\Delta_{\psi_T}(w^*, w_{T+1}) + \frac{1}{\eta} \sum_{t=1}^{T-1} \Delta_{\psi_{t+1}}(w^*, w_t) - \Delta_{\psi_t}(w^*, w_t) \\
&\le \frac{1}{\eta}\Delta_{\psi_1}(w^*, w_1) + \underbrace{\frac{1}{\eta} \sum_{t=1}^{T-1} \Delta_{\psi_{t+1}}(w^*, w_t) - \Delta_{\psi_t}(w^*, w_t)}_{T_3} .
\end{aligned}$$

Next, we use $\Delta_\psi(w^*, w_t) = \frac{1}{2}\|w^* - w_t\|_{H_t}^2 = \langle w^* - w_t, H_t(w^* - w_t)\rangle$ to write

$$\begin{aligned} T_3 &= \frac{1}{\eta}\sum_{t=1}^{T-1}\langle w^* - w_t, (H_{t+1} - H_t)(w^* - w_t)\rangle \\ &\leq \frac{1}{\eta}\sum_{t=1}^{T-1}\|w^* - w_t\|_\infty \cdot \|s_{t+1} - s_t\|_1 \end{aligned}$$

Since each component of s_t is non-negative, $\|s_{t+1} - s_t\|_1 = \langle s_{t+1} - s_t, \mathbf{1}\rangle$. Moreover, if we let $F_\infty = \max_t \|w^* - w_t\|_\infty$ then we have

$$T_3 \leq \frac{1}{\eta}F_\infty \cdot \langle s_T, \mathbf{1}\rangle - \frac{1}{\eta}\|w^* - w_1\|\langle s_1, \mathbf{1}\rangle\,.$$

Using $\Delta_{\psi_1}(w^*, w_1) \leq \|w^* - w_1\|\langle s_1, \mathbf{1}\rangle$, we can conclude that

$$T_1 \leq \frac{1}{\eta}F_\infty \cdot \langle s_T, \mathbf{1}\rangle\,.$$

In order to bound T_2 we need the following lemma [MS10]:

Lemma 1.6 *For any non-negative real numbers $x_1, \ldots, x_T$ we have*

$$\frac{1}{2}\sum_{t=1}^{T}\frac{x_t}{\sqrt{\sum_{t'=1}^{t} x_{t'}}} \leq \sqrt{\sum_{t=1}^{T} x_t} \tag{1.11}$$

Proof By induction, suppose (1.11) holds for $T-1$, then

$$\begin{aligned} \frac{1}{2}\sum_{t=1}^{T}\frac{x_t}{\sqrt{\sum_{t'=1}^{t} x_{t'}}} &\leq \sqrt{\sum_{t=1}^{T-1} x_t} + \frac{1}{2}\frac{x_T}{\sqrt{\sum_{t'=1}^{T} x_{t'}}} \\ &= \sqrt{Z - x} + \frac{1}{2}\frac{x}{\sqrt{Z}} \end{aligned}$$

where $Z = \sum_{t=1}^{T} x_t$ and $x = x_T$. Clearly the right hand size is a concave function of x. The gradient with respect to x is given by $\frac{-1}{2\sqrt{Z-x}} + \frac{1}{2\sqrt{Z}}$ which is negative for $x \geq 0$. This shows that, subject to the constraint $x \geq 0$, the right hand size is maximized at $x = 0$, in which case its value is $\sqrt{Z}$. ■

Applying the above lemma to the sequences $g_{1,j}, \ldots, g_{T,j}$ for $j = 1, \ldots, d$ shows that

$$T_2 \leq \eta\langle s_T, \mathbf{1}\rangle$$

Algorithm 1.3 Stochastic Mirror Descent for Composite Functions

```
Input: Initial point $w_1$, maximum iterations $T$
for $t = 1, \ldots, T$ do
    Compute $\hat{w}_{t+1} = \mathrm{argmin}_w \, \eta_t \langle g_t, w \rangle + \eta r(w) + \Delta_\psi(w, w_t)$ with $g_t := \partial_w f_t(w_t)$
    Set $w_{t+1} = P_{\psi,\Omega}(\hat{w}_{t+1})$
end for
Return: $w_{T+1}$
```

Combining the bounds on T_1 and T_2 obtains the following expression for the regret

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w^*) \leq \frac{1}{\eta} F_\infty \cdot \langle s_T, \mathbf{1} \rangle + \eta \langle s_T, \mathbf{1} \rangle$$

1.3.2 Dealing with Composite Objective Functions

Next we consider algorithms for solving the following so-called composite problem:

$$\min_{w \in \Omega} J(w) + r(w) \text{ where } J(w) = \sum_{t=1}^{T} f_t(w), \tag{1.12}$$

and $r(w)$ is a simple to evaluate regularizer. For instance, $r(w) = \|w\|^2$ or $r(w) = \|w\|_1^2$ etc. We will operate under the same assumptions as in the previous sub-section. The algorithm that we will employ is given in Algorithm 1.3. Note that Algorithm 1.2 is recovered as a special case when $r(w) = 0$. Now we prove the analog of Lemma 1.2 for composite functions.

Lemma 1.7 *Let f_t be strongly convex with respect to ψ with modulus $\lambda \geq 0$ for all t. For any $w \in \Omega$ the sequences generated by Algorithm 1.2 satisfy*

$$\Delta_\psi(w, w_{t+1}) \leq \Delta_\psi(w, w_t) - \eta_t \langle g_t, w_t - w \rangle - \eta_t \langle \nabla r(w_{t+1}), w_{t+1} - w \rangle + \frac{\eta_t^2}{2\sigma} \|g_t\|^2 \tag{1.13}$$

$$\leq (1 - \eta_t \lambda) \Delta_\psi(w, w_t) - \eta_t (f_t(w_t) - f_t(w)) - \eta_t (r(w_{t+1}) - r(w)) + \frac{\eta_t^2}{2\sigma} \|g_t\|^2 . \tag{1.14}$$

Proof We prove the result in three steps. First we upper bound $\Delta_\psi(w, w_{t+1})$ by $\Delta_\psi(w, \hat{w}_{t+1})$. This is a consequence of (**??**) and the non-negativity of the

Bregman divergence which allows us to write

$$\Delta_\psi(w, w_{t+1}) \leq \Delta_\psi(w, \hat{w}_{t+1}). \tag{1.15}$$

In the next step we use Lemma **??** to write

$$\Delta_\psi(w, w_t) + \Delta_\psi(w_t, \hat{w}_{t+1}) - \Delta_\psi(w, \hat{w}_{t+1}) = \langle \nabla\psi(\hat{w}_{t+1}) - \nabla\psi(w_t), w - w_t \rangle .$$

Since $\nabla\psi^* = (\nabla\psi)^{-1}$, the update in step 3 of Algorithm 1.3 can equivalently be written as $\nabla\psi(\hat{w}_{t+1}) - \nabla\psi(w_t) = -\eta_t g_t - \eta_t \nabla r(w_{t+1})$. Plugging this in the above equation and rearranging

$$\Delta_\psi(w, \hat{w}_{t+1}) = \Delta_\psi(w, w_t) - \eta_t \langle g_t, w_t - w \rangle - \eta_t \langle \nabla r(w_{t+1}), w_t - w \rangle + \Delta_\psi(w_t, \hat{w}_{t+1}). \tag{1.16}$$

Finally we upper bound $\Delta_\psi(w_t, \hat{w}_{t+1})$. For this we need two observations: First, $\langle x, y \rangle \leq \frac{1}{2\sigma} \|x\|^2 + \frac{\sigma}{2} \|y\|^2$ for all $x, y \in \mathbb{R}^n$ and $\sigma > 0$. Second, the σ strong convexity of ψ allows us to bound $\Delta_\psi(\hat{w}_{t+1}, w_t) \geq \frac{\sigma}{2} \|w_t - \hat{w}_{t+1}\|^2$. Using these two observations

$$\begin{aligned}
\Delta_\psi(w_t, \hat{w}_{t+1}) &= \psi(w_t) - \psi(\hat{w}_{t+1}) - \langle \nabla\psi(\hat{w}_{t+1}), w_t - \hat{w}_{t+1} \rangle \\
&= -(\psi(\hat{w}_{t+1}) - \psi(w_t) - \langle \nabla\psi(w_t), \hat{w}_{t+1} - w_t \rangle) \\
&\quad + \langle \eta_t g_t, w_t - \hat{w}_{t+1} \rangle + \eta_t \langle \nabla r(w_{t+1}), w_t - \hat{w}_{t+1} \rangle \\
&= -\Delta_\psi(\hat{w}_{t+1}, w_t) + \langle \eta_t g_t, w_t - \hat{w}_{t+1} \rangle + \eta_t \langle \nabla r(w_{t+1}), w_t - \hat{w}_{t+1} \rangle \\
&\leq -\frac{\sigma}{2} \|w_t - \hat{w}_{t+1}\|^2 + \frac{\eta_t^2}{2\sigma} \|g_t\|^2 + \frac{\sigma}{2} \|w_t - \hat{w}_{t+1}\|^2 + \eta_t \langle \nabla r(w_{t+1}), w_t - \hat{w}_{t+1} \rangle \\
&= \frac{\eta_t^2}{2\sigma} \|g_t\|^2 + \eta_t \langle \nabla r(w_{t+1}), w_t - \hat{w}_{t+1} \rangle . \qquad (1.17)
\end{aligned}$$

Inequality (1.13) follows by putting together (1.15), (1.16), (1.17), and simplifying while (1.14) follows by using (**??**) with $f = f_t$ and $w' = w_t$ and substituting into (1.13). ■

Problems

Problem 1.1 (Generalized Cauchy-Schwartz {1}) *Show that* $\langle x, y \rangle \leq \frac{1}{2\sigma} \|x\|^2 + \frac{\sigma}{2} \|y\|^2$ *for all* $x, y \in \mathbb{R}^n$ *and* $\sigma > 0$.

Problem 1.2 (Bounding sum of a series {1}) *Show that* $\sum_{t=a}^{b} \frac{1}{2\sqrt{t}} \leq \sqrt{b - a + 1}$. ***Hint:*** *Upper bound the sum by an integral.*

Bibliography

[BT03] Amir Beck and Marc Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters **31** (2003), no. 3, 167–175.

[MS10] H. Brendan McMahan and Matthew J. Streeter, *Adaptive bound optimization for online convex optimization.*, COLT (Adam Tauman Kalai and Mehryar Mohri, eds.), Omnipress, 2010, pp. 244–256.

[Ned02] Angelia Nedić, *Subgradient methods for convex minimization*, Ph.D. thesis, MIT, 2002.

[SSS07] S. Shalev-Shwartz and Y. Singer, *Logarithmic regret algorithms for strongly convex repeated games*, Tech. report, School of Computer Science, Hebrew University, 2007.

[Zin03] M. Zinkevich, *Online convex programming and generalised infinitesimal gradient ascent*, Proceedings of the International Conference on Machine Learning, 2003, pp. 928–936.