Supplementary: Learning for Structured Prediction
Using Approximate Subgradient Descent with Working Sets

Aurélien Lucchi\textsuperscript{1}∗ Yunpeng Li\textsuperscript{1} Pascal Fua\textsuperscript{1}
\textsuperscript{1} Computer Vision Laboratory, EPFL, Lausanne

We analyze the convergence properties of Algorithm 1. Recall that our goal is to minimize the empirical objective function:

\[ L(w) = \frac{1}{N} \sum_{n=1}^{N} l(Y^n, Y^*, w) + \frac{1}{2C} ||w||^2. \]  

(1)

At each iteration Algorithm 1 chooses a random training example \((X^n, Y^n)\) by picking an index \(n \in \{1...N\}\) uniformly at random. We then replace the objective in Eq. 1 with an approximation based on the training example \((X^n, Y^n)\), yielding:

\[ f(w, n) = l(Y^n, Y^*, w) + \frac{1}{2C} ||w||^2. \]

(2)

We consider the case where \(l : \mathcal{W} \rightarrow \mathbb{R}\) is a convex loss function so that \(f(w)\) is a \(\lambda\)-strongly convex function where \(\lambda = \frac{1}{C}\).

We assume we can compute an \(\epsilon\)-subgradient of \(f(w)\) whose magnitude is bounded by a constant \(G\), i.e. \(||g||^2 \leq G^2\). It can be shown that such function is \(G\)-Lipschitz continuous such that:

\[ \forall w' \in \mathcal{W}, g^T(w - w') \geq f(w) - f(w') - \epsilon \]

(3)

Let \(w^*\) be the minimizer of \(L(w)\). The following relation then holds trivially for \(w^*\):

\[ g^T(w - w^*) \geq f(w) - f(w^*) - \epsilon \]

(4)

1. Convergence properties of the t-th parameter vector

1.1. Proof of convergence

This proof for subgradients was derived in [1] and we extend it to approximate subgradients here. We first present some inequalities that will be used in the following proof.

By the strong convexity of \(f(w)\), we have:

\[ \langle g^{(t)}, w^{(t)} - w^* \rangle \geq f(w^{(t)}) - f(w^*) + \frac{\lambda}{2} ||w^{(t)} - w^*||^2 - \epsilon \]

(5)

Because \(w^*\) minimizes \(f(w)\), we have:

\[ f(w^{(t)}) - f(w^*) \geq \frac{\lambda}{2} ||w^{(t)} - w^*||^2 \]

(6)

By combining Eq. 5 and 6 we get:

\[ \langle g^{(t)}, w^{(t)} - w^* \rangle \geq \lambda ||w^{(t)} - w^*||^2 - \epsilon \]

(7)

In the following, we first start by bounding \(||w^{(1)} - w^*||\) and then derive a bound for \(\mathbb{E}||w^{(t+1)} - w^*||\).

∗This work was supported in part by the EU ERC Grant MicroNano.
**Lemma 1.** The error of $w^{(1)}$ is:

$$\|w^{(1)} - w^*\|_2^2 \leq \frac{G^2 + 2\epsilon \lambda}{\lambda^2} \quad (8)$$

**Proof.** From Eq. 5, we deduce:

$$\langle g^{(1)}, w^{(1)} - w^* \rangle \geq f(w^{(1)}) - f(w^*) + \frac{\lambda}{2} \|w^{(1)} - w^*\|_2^2 - \epsilon$$

$$\geq \frac{\lambda}{2} \|w^{(1)} - w^*\|_2^2 + \frac{\lambda}{2} \|w^{(1)} - w^*\|_2^2 - \epsilon$$

$$\geq \lambda \|w^{(1)} - w^*\|_2^2 - \epsilon,$$  \quad (9)

where the last inequality follows from the fact that $f(w^{(1)}) - f(w^*) \geq 0$. Using the Cauchy-Schwarz inequality ($\langle X, Y \rangle \leq \|X\|\|Y\|$), we get:

$$\|g^{(1)}\|_2^2 \geq (\lambda \|w^{(1)} - w^*\|_2^2 - \epsilon) / \|w^{(1)} - w^*\|_2^2$$

$$= \lambda^2 \|w^{(1)} - w^*\|_2^2 - 2\epsilon \lambda + \frac{\epsilon^2}{\|w^{(1)} - w^*\|_2^2} \quad (10)$$

Using the assumption that $\|g^{(t)}\|_2^2 \leq G^2$, we get:

$$G^2 \geq \lambda^2 \|w^{(1)} - w^*\|_2^2 - 2\epsilon \lambda + \frac{\epsilon^2}{\|w^{(1)} - w^*\|_2^2} \quad (11)$$

We then derive the following bound for $\|w^{(1)} - w^*\|_2^2$:

$$\|w^{(1)} - w^*\|_2^2 \leq \max \left( \frac{G^2 + 2\epsilon \lambda}{\lambda^2}, - \frac{\epsilon^2}{G^2 + 2\epsilon \lambda} \right)$$  \quad (12)

$$\frac{G^2 + 2\epsilon \lambda}{\lambda^2} - \frac{\epsilon^2}{G^2 + 2\epsilon \lambda} = \frac{(G^2 + 2\epsilon \lambda)(G^2 + 2\epsilon \lambda) - \epsilon^2 \lambda^2}{\lambda^2(G^2 + 2\epsilon \lambda)}$$

$$= \frac{(G^2 + 2\epsilon \lambda + \epsilon \lambda)(G^2 + 2\epsilon \lambda - \epsilon \lambda)}{\lambda^2(G^2 + 2\epsilon \lambda)} = \frac{(G^2 + 2\epsilon \lambda)(G^2 + 2\epsilon \lambda)}{\lambda^2(G^2 + 2\epsilon \lambda)} \geq 0 \quad (13)$$

We get Eq. 8 by combining Eq. 12 and 13.

\[\square\]

**Theorem 1.** The error of $w^{(t+1)}$ is:

$$\mathbb{E}\|w^{(t+1)} - w^*\|_2^2 \leq \frac{G^2}{\lambda^2} + \epsilon \quad (14)$$

**Proof.**

$$\mathbb{E}\|w^{(t+1)} - w^*\|_2^2 = \mathbb{E}\|w^{(t)} - \eta^{(t)} g^{(t)} - w^*\|_2^2$$

$$= \mathbb{E}\|w^{(t)} - w^*\|_2^2 - 2\eta^{(t)} \mathbb{E}(\langle g^{(t)}, (w^{(t)} - w^*) \rangle) + (\eta^{(t)})^2 \mathbb{E}\|g^{(t)}\|_2^2$$

$$\leq \mathbb{E}\|w^{(t)} - w^*\|_2^2 - 2\eta^{(t)} (\lambda \mathbb{E}\|w^{(t)} - w^*\|_2^2 - \epsilon) + (\eta^{(t)})^2 G^2$$

$$= (1 - 2\eta^{(t)} \lambda) \mathbb{E}\|w^{(t)} - w^*\|_2^2 + (\eta^{(t)})^2 G^2 + 2\eta^{(t)} \epsilon \quad (15)$$

By applying the inequality recursively:
\[
\mathbb{E}\|w^{(t+1)} - w^*\|^2_2 \leq (1 - 2\eta^{(t)}\lambda)\mathbb{E}\|w^{(t)} - w^*\|^2_2 + \eta^{(t)}G^2 + 2\eta^{(t)}\epsilon
\]
\[
\leq (1 - 2\eta^{(t)}\lambda)(1 - 2\eta^{(t-1)}\lambda)\mathbb{E}\|w^{(t-1)} - w^*\|^2_2 + (\eta^{(t-1)})^2 G^2 + 2\eta^{(t-1)}\epsilon + (\eta^{(t)})^2 G^2 + 2\eta^{(t)}\epsilon
\]
\[
\leq \prod_{i=2}^t (1 - 2\eta^{(i)}\lambda)(\mathbb{E}\|w^{(2)} - w^*\|^2_2) + \sum_{i=2}^t \prod_{j=i+1}^t (1 - 2\eta^{(j)}\lambda)(\eta^{(j)})^2 G^2 + \sum_{i=2}^t \prod_{j=i+1}^t (1 - 2\eta^{(j)}\lambda)2\eta^{(j)}\epsilon
\]
\[\sum_{i=2}^t \prod_{j=i+1}^t (1 - 2\eta^{(j)}\lambda)2\eta^{(j)}\epsilon \tag{16}\]

Plugging in \(\eta^{(i)}\), we get:
\[
\mathbb{E}\|w^{(t+1)} - w^*\|^2_2 \leq \prod_{i=2}^t \left(1 - \frac{2}{i}\right) \left(\mathbb{E}\|w^{2)} - w^*\|^2_2\right) + \sum_{i=2}^t \prod_{j=i+1}^t \left(1 - \frac{2}{j}\right) \left(\frac{1}{i}\right)^2 \frac{G^2}{\lambda^2} + \sum_{i=2}^t \prod_{j=i+1}^t \left(1 - \frac{2}{j}\right) \frac{2\epsilon}{i} \tag{17}\]

Rakhlin [1] showed that setting \(\eta^{(i)} = \frac{1}{\lambda t}\) gives us a \(O(1/t)\) rate. Indeed, we have:
\[
\prod_{j=i+1}^t \left(1 - \frac{2}{j}\right) = \prod_{j=i+1}^t \left(\frac{j-2}{j}\right) = \frac{(i-1)i}{(t-1)t} \tag{18}\]
and therefore
\[
\sum_{i=2}^t \frac{1}{i^2} \prod_{j=i+1}^t \left(1 - \frac{2}{j}\right) = \sum_{i=2}^t \frac{(i-1)i}{i(t-1)t} \leq \frac{1}{t} \tag{19}\]
\[
\sum_{i=2}^t \prod_{j=i+1}^t \left(1 - \frac{2}{j}\right) \frac{2\epsilon}{i} = \sum_{i=2}^t \frac{2(i-1)i\epsilon}{i(t-1)t} = \frac{2\epsilon}{(t-1)t} \sum_{i=1}^{t-1} i = \frac{2\epsilon}{(t-1)t} \frac{(t-1)t}{2} = \epsilon \tag{20}\]

By combining Eq. 17 with Eq. 19 and Eq. 20, we then get:
\[
\mathbb{E}\|w^{(t+1)} - w^*\|^2_2 \leq \frac{G^2}{\lambda^2} + \epsilon \tag{21}\]

Given that \(L\) is a \(\lambda\)-strongly convex function, the following bound can be derived from Eq. 6:
\[
\mathbb{E}\|L(w^{(t+1)}) - L(w^*)\|^2_2 \leq \frac{G^2}{\lambda t} + \frac{\epsilon}{\lambda} \tag{22}\]

We can deduce that the conditions of convergence are the same as the ones for subgradient descent (i.e. for \(\epsilon = 0\)):
\[
\lim_{T \to +\infty} \sum_{i=1}^T \eta^{(i)} \to \infty \tag{23}\]
\[
\lim_{T \to +\infty} \sum_{i=1}^T (\eta^{(i)})^2 \to 0 \tag{23}\]

As long as the choice of the step size satisfies Eq. 23, we can see that the first term on the right side of Eq. 21 and Eq. 22 goes to 0 so stochastic \(\epsilon\)-subgradient descent will convergence to a distance \(\epsilon\) away from the optimal value.

\[\Box\]
References