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

Aurélien Lucchi[1*] Yunpeng Li[1] Pascal Fua[1]
[1] Computer Vision Laboratory, EPFL, Lausanne

We analyze the convergence properties of Algorithm 1. Recall that our goal is to find the parameter vector $\mathbf{w}^*$ that minimizes the empirical objective function:

$$\mathcal{L}(\mathbf{w}) = \sum_{n=1}^{N} l(Y^n, Y^*, \mathbf{w}) + \frac{1}{2C} ||\mathbf{w}||^2 . \tag{1}$$

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

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

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

Recall that the definition of an ϵ-subgradient of $f(\mathbf{w})$ is:

$$\forall \mathbf{w}' \in \mathcal{W}, \mathbf{g}^T(\mathbf{w} - \mathbf{w}') \geq f(\mathbf{w}) - f(\mathbf{w}') - \epsilon. \tag{3}$$

In the following, we will assume that the magnitude of the ϵ-subgradients we compute is bounded by a constant G, i.e. $||g||_2^2 \leq G^2$.

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

$$\mathbf{g}^T(\mathbf{w} - \mathbf{w}^*) \geq f(\mathbf{w}) - f(\mathbf{w}^*) - \epsilon. \tag{4}$$

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 $f(\mathbf{w})$, we have:

$$\langle g^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^* \rangle \geq f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*) + \frac{\lambda}{2} \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - \epsilon. \tag{5}$$

An equivalent condition is:

$$\langle g^{(t)}, \mathbf{w}^{(t)} - \mathbf{w}^* \rangle \geq \lambda \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - \epsilon. \tag{6}$$

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

Lemma 1. *The error of* $\mathbf{w}^{(1)}$ *is:*

$$\|\mathbf{w}^{(1)} - \mathbf{w}^*\|_2^2 \leq \frac{G^2 + 2\epsilon\lambda}{\lambda^2}. \tag{7}$$

*This work was supported in part by the EU ERC Grant MicroNano.

Proof. From Eq. 6, we have:

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

Using the Cauchy-Schwarz inequality ($|\langle X, Y \rangle| \leq \|X\|\|Y\|$), we get:

$$\begin{aligned} \|g^{(1)}\|_2^2 &\geq \frac{\left(\lambda \|\mathbf{w}^{(1)} - \mathbf{w}^*\|_2^2 - \epsilon\right)^2}{\|\mathbf{w}^{(1)} - \mathbf{w}^*\|_2^2} \\ &= \lambda^2 \|\mathbf{w}^{(1)} - \mathbf{w}^*\|_2^2 - 2\epsilon\lambda + \frac{\epsilon^2}{\|\mathbf{w}^{(1)} - \mathbf{w}^*\|_2^2}, \end{aligned} \tag{8}$$

and from the assumption that $\|g^{(t)}\|^2 \leq G^2$, we have that:

$$G^2 \geq \lambda^2 \|\mathbf{w}^{(1)} - \mathbf{w}^*\|_2^2 - 2\epsilon\lambda + \frac{\epsilon^2}{\|\mathbf{w}^{(1)} - \mathbf{w}^*\|_2^2}. \tag{9}$$

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

$$\|\mathbf{w}^{(1)} - \mathbf{w}^*\|_2^2 \leq \max\left(\frac{G^2 + 2\epsilon\lambda}{\lambda^2}, \frac{\epsilon^2}{G^2 + 2\epsilon\lambda}\right). \tag{10}$$

$$\begin{aligned} \frac{G^2 + 2\epsilon\lambda}{\lambda^2} - \frac{\epsilon^2}{G^2 + 2\epsilon\lambda} &= \frac{(G^2 + 2\epsilon\lambda)(G^2 + 2\epsilon\lambda) - \epsilon^2\lambda^2}{\lambda^2(G^2 + 2\epsilon\lambda)} = \frac{(G^2 + 2\epsilon\lambda)^2 - \epsilon^2\lambda^2}{\lambda^2(G^2 + 2\epsilon\lambda)} \\ &= \frac{(G^2 + 2\epsilon\lambda + \epsilon\lambda)(G^2 + 2\epsilon\lambda - \epsilon\lambda)}{\lambda^2(G^2 + 2\epsilon\lambda)} = \frac{(G^2 + 3\epsilon\lambda)(G^2 + \epsilon\lambda)}{\lambda^2(G^2 + 2\epsilon\lambda)} \geq 0. \end{aligned} \tag{11}$$

Therefore, we see that:

$$\max\left(\frac{G^2 + 2\epsilon\lambda}{\lambda^2}, \frac{\epsilon^2}{G^2 + 2\epsilon\lambda}\right) = \frac{G^2 + 2\epsilon\lambda}{\lambda^2}. \tag{12}$$

We get Eq. 7 by combining Eq. 10 and 12 .

□

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

$$\mathbb{E}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq \frac{G^2}{\lambda^2 t} + \frac{\epsilon}{\lambda}. \tag{13}$$

Proof.

$$\begin{aligned} \mathbb{E}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 &= \mathbb{E}\|\mathbf{w}^{(t)} - \eta^{(t)}\mathbf{g}^{(t)} - \mathbf{w}^*\|_2^2 \\ &= \mathbb{E}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta^{(t)}\mathbb{E}(\langle \mathbf{g}^{(t)}, (\mathbf{w}^{(t)} - \mathbf{w}^*) \rangle) + (\eta^{(t)})^2(\mathbb{E}\|\mathbf{g}^{(t)}\|_2^2) \\ &\leq \mathbb{E}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta^{(t)}(\lambda\mathbb{E}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - \epsilon) + (\eta^{(t)})^2 G^2 \\ &= (1 - 2\eta^{(t)}\lambda)\mathbb{E}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + (\eta^{(t)})^2 G^2 + 2\eta^{(t)}\epsilon \end{aligned} \tag{14}$$

By applying the inequality recursively:

$$\begin{aligned} \mathbb{E}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 &\leq (1 - 2\eta^{(t)}\lambda)\mathbb{E}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + (\eta^{(t)})^2 G^2 + 2\eta^{(t)}\epsilon \\ &\leq (1 - 2\eta^{(t)}\lambda)((1 - 2\eta^{(t-1)}\lambda)\mathbb{E}\|\mathbf{w}^{(t-1)} - \mathbf{w}^*\|_2^2 + (\eta^{(t-1)})^2 G^2 + 2\eta^{(t-1)}\epsilon) + (\eta^{(t)})^2 G^2 + 2\eta^{(t)}\epsilon \\ &\leq \left(\prod_{i=2}^{t}(1 - 2\eta^{(i)}\lambda)\right)(\mathbb{E}\|\mathbf{w}^{(2)} - \mathbf{w}^*\|_2^2) + \sum_{i=2}^{t}\prod_{j=i+1}^{t}(1 - 2\eta^{(j)}\lambda)(\eta^{(i)})^2 G^2 + \\ &\quad \sum_{i=2}^{t}\prod_{j=i+1}^{t}(1 - 2\eta^{(j)}\lambda)2\eta^{(i)}\epsilon. \end{aligned} \tag{15}$$

Plugging in $\eta^{(i)} = \frac{1}{\lambda i}$, we get:

$$\begin{aligned}\mathbb{E}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 &\leq \prod_{i=2}^{t}\left(1-\frac{2}{i}\right)(\mathbb{E}\|\mathbf{w}^{(2)} - \mathbf{w}^*\|_2^2) + \sum_{i=2}^{t}\prod_{j=i+1}^{t}\left(1-\frac{2}{j}\right)\left(\frac{1}{i}\right)^2\frac{G^2}{\lambda^2} \\ &+ \sum_{i=2}^{t}\prod_{j=i+1}^{t}\left(1-\frac{2}{j}\right)\frac{2\epsilon}{i\lambda} \\ &= \frac{G^2}{\lambda^2}\sum_{i=2}^{t}\prod_{j=i+1}^{t}\left(1-\frac{2}{j}\right)\left(\frac{1}{i}\right)^2 + \sum_{i=2}^{t}\prod_{j=i+1}^{t}\left(1-\frac{2}{j}\right)\frac{2\epsilon}{i\lambda} \end{aligned} \tag{16}$$

Rakhlin [1] showed that setting $\eta^{(i)} = \frac{1}{\lambda i}$ gives us a $O(1/t)$ rate. Indeed, we have:

$$\prod_{j=i+1}^{t}\left(1-\frac{2}{j}\right) = \prod_{j=i+1}^{t}\left(\frac{j-2}{j}\right) = \frac{(i-1)i}{(t-1)t}, \tag{17}$$

and therefore

$$\sum_{i=2}^{t}\frac{1}{i^2}\prod_{j=i+1}^{t}\left(1-\frac{2}{j}\right) = \sum_{i=2}^{t}\frac{(i-1)}{i(t-1)t} \leq \frac{1}{t}, \tag{18}$$

$$\sum_{i=2}^{t}\prod_{j=i+1}^{t}\left(1-\frac{2}{j}\right)\frac{2\epsilon}{i\lambda} = \sum_{i=2}^{t}\frac{2(i-1)i\epsilon}{i(t-1)t\lambda} = \frac{2\epsilon}{(t-1)t\lambda}\sum_{i=1}^{t-1} i = \frac{2\epsilon}{(t-1)t\lambda}\left(\frac{(t-1)t}{2}\right) = \frac{\epsilon}{\lambda} \tag{19}$$

By combining Eq. 16 with Eq. 18 and Eq. 19, we then get:

$$\mathbb{E}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq \frac{G^2}{\lambda^2 t} + \frac{\epsilon}{\lambda}. \tag{20}$$

We can deduce that the conditions of convergence are the same as the ones for subgradient descent (i.e. for $\epsilon = 0$) :

$$\begin{aligned}\lim_{T\to+\infty}\sum_{i=1}^{T}\eta^{(i)} \to \infty \\ \lim_{T\to+\infty}\sum_{i=1}^{T}(\eta^{(i)})^2 < \infty\end{aligned} \tag{21}$$

As long as the choice of the step size satisfies Eq. 21, we can see that the first term on the right side of Eq. 20 goes to 0 so stochastic ϵ-subgradient descent will convergence to a distance $\frac{\epsilon}{\lambda}$ away from the optimal value.

□

References

[1] A. Rakhlin, O. Shamir, and K. Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. Technical report, ArXiv, 2012. 1, 3