Learning Discriminative Features and Structured Models for Segmentation of 2D and 3D Imagery

Aurélien Lucchi[1] Pablo Márquez Neila[2] Carlos Becker[2]
Yunpeng Li[2] Kevin Smith[3]
Graham Knott[4] Pascal Fua[2]
[1] Department of Computer Science, ETHZ, Zürich, Switzerland
[2] Computer Vision Laboratory, EPFL, Lausanne, Switzerland
[3]Biozentrum, University of Basel, Switzerland
[4] Interdisciplinary Center for Electron Microscopy, EPFL, Lausanne, Switzerland

APPENDIX

A. Convergence for decreasing step size

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 tth parameter vector: 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:

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

An equivalent condition is:

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

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:

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

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

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

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:

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

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

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

and therefore

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

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

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

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

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

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

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.

□

B. Convergence for adaptive step size

We start by deriving a bound on the adaptive step size $\eta^{(t)}$ before giving a proof of convergence of Algorithm 2.

1) Bound on the step-size: Recall that the step size $\eta^{(t)}$ is chosen to minimize the following heuristic function:

$$\begin{aligned}
\eta^{(t)} \leftarrow &\quad \arg\min_{\eta} h^R_{\mathcal{A}^n,\mathbf{w}^{(t)}}(\eta) \\
= &\quad \arg\min_{\eta}\max_{Y\in\mathcal{A}^n} \rho^{(t)}\left|\left|\mathbf{g}^{(t)}\right|\right|^2\eta^2 + l\left(Y^n, Y, \mathbf{w}^{(t)} - \eta\mathbf{g}^{(t)}\right) + \frac{\lambda}{2}\left|\left|\mathbf{w}^{(t)} - \eta\mathbf{g}^{(t)}\right|\right|. \qquad (22)
\end{aligned}$$

We will make use of the following definitions from the paper:

- $\rho^{(t)}$ is a constant that controls the regularization strength. We choose $\rho^{(t)} = Ct$ where C is a constant value that can be determined by cross-validation
- $\mathbf{g}^{(t)}$ is a subgradient of $f(Y^n, Y^*_{\mathbf{w}}, \mathbf{w}) = l(Y^n, Y^*_{\mathbf{w}}, \mathbf{w}) + \frac{\lambda}{2}$
- λ is a constant that controls the trade-off between margin and training error

Using the definition of the hinge loss function l, the second term in Eq. 22 becomes:

$$\begin{aligned}
l\left(Y^n, Y, \mathbf{w}^{(t)} - \eta\mathbf{g}^{(t)}\right) &= \left\langle \mathbf{w}^{(t)} - \eta\mathbf{g}^{(t)}, \psi(X,Y)\right\rangle + \Delta(Y^n, Y) - \left\langle \mathbf{w}^{(t)} - \eta\mathbf{g}^{(t)}, \psi(X,Y^n)\right\rangle \\
&= \left\langle \mathbf{w}^{(t)}, \psi(X,Y) - \psi(X,Y^n)\right\rangle - \eta\left\langle \mathbf{g}^{(t)}, \psi(X,Y) - \psi(X,Y^n)\right\rangle + \Delta(Y^n, Y) \\
&= \left\langle \mathbf{w}^{(t)}, \delta\psi_n(Y)\right\rangle - \eta\left\langle \mathbf{g}^{(t)}, \delta\psi_n(Y)\right\rangle + \Delta(Y^n, Y), \qquad (23)
\end{aligned}$$

where $\delta\psi_n(Y) = \psi(X,Y) - \psi(X,Y^n)$.

The last term in Eq. 22 is:

$$\begin{aligned}\frac{\lambda}{2}\left|\left|\mathbf{w}^{(t)} - \eta\mathbf{g}^{(t)}\right|\right|^2 &= \frac{\lambda}{2}\left(\left|\left|\mathbf{w}^{(t)}\right|\right|^2 - 2\left\langle\mathbf{w}^{(t)}, \eta\mathbf{g}^{(t)}\right\rangle + \left|\left|\eta\mathbf{g}^{(t)}\right|\right|^2\right) \\ &= \frac{\lambda}{2}\left|\left|\mathbf{w}^{(t)}\right|\right|^2 + \left(\frac{\lambda}{2}\right)\eta^2\left|\left|\mathbf{g}^{(t)}\right|\right|^2 - \eta\left\langle\lambda\mathbf{w}^{(t)}, \mathbf{g}^{(t)}\right\rangle. \end{aligned} \tag{24}$$

By the definition of $\eta^{(t)}$ in Eq. 22, we know that $h_{\mathcal{A}^n,\mathbf{w}^{(t)}}(\eta^{(t)}) \leq h_{\mathcal{A}^n,\mathbf{w}^{(t)}}(0)$. Combining Eq. 23 and 24 into Eq. 22, we get:

$$\begin{aligned}\max_{Y\in\mathcal{A}_+^n}&\left(\frac{\lambda}{2} + \rho^{(t)}\right)\left|\left|\mathbf{g}^{(t)}\right|\right|^2(\eta^{(t)})^2 - \left\langle\mathbf{g}^{(t)}, \delta\psi_n(Y) + \lambda\mathbf{w}^{(t)}\right\rangle\eta^{(t)} \\ &+ \left(\frac{\lambda}{2}\left|\left|\mathbf{w}^{(t)}\right|\right|^2 + \left\langle\mathbf{w}^{(t)}, \delta\psi_n(Y)\right\rangle + \Delta(Y^n, Y)\right) \\ &\leq \max_{Y\in\mathcal{A}_+^n}\left(\frac{\lambda}{2}\left|\left|\mathbf{w}^{(t)}\right|\right|^2 + \left\langle\mathbf{w}^{(t)}, \delta\psi_n(Y)\right\rangle + \Delta(Y^n, Y)\right)\end{aligned} \tag{25}$$

Canceling the constant terms that do not depend on $\eta^{(t)}$ on both sides, and using the shorthand notation $Y^* = \arg\max_{Y\in\mathcal{A}_+^n} h_{\mathcal{A}^n,\mathbf{w}^{(t)}}(0)$, we get:

$$\left(\frac{\lambda}{2} + \rho^{(t)}\right)\left|\left|\mathbf{g}^{(t)}\right|\right|^2(\eta^{(t)})^2 - \left\langle\mathbf{g}^{(t)}, \delta\psi_n(Y^*) + \lambda\mathbf{w}^{(t)}\right\rangle\eta^{(t)} \leq 0 \tag{26}$$

$$\eta^{(t)}\underbrace{\left(\left(\frac{\lambda}{2} + \rho^{(t)}\right)\left|\left|\mathbf{g}^{(t)}\right|\right|^2\eta^{(t)} - \left\langle\mathbf{g}^{(t)}, \delta\psi_n(Y^*) + \lambda\mathbf{w}^{(t)}\right\rangle\right)}_{=T(\eta^{(t)})} \leq 0. \tag{27}$$

In order to solve Eq. 27, we consider separately the following two cases: either $\eta^{(t)} \geq 0$ and $T(\eta^{(t)}) \leq 0$, or $\eta^{(t)} \leq 0$ and $T(\eta^{(t)}) \geq 0$.

In the first case, we have:

$$\left(\frac{\lambda}{2} + \rho^{(t)}\right)\left|\left|\mathbf{g}^{(t)}\right|\right|^2\eta^{(t)} - \left\langle\mathbf{g}^{(t)}, \delta\psi_n(Y^*) + \lambda\mathbf{w}^{(t)}\right\rangle \leq 0 \tag{28}$$

$$\left(\frac{\lambda}{2} + \rho^{(t)}\right)\left|\left|\delta\psi_n(Y^*) + \lambda\mathbf{w}^{(t)}\right|\right|^2\eta^{(t)} - \left\langle\delta\psi_n(Y^*) + \lambda\mathbf{w}^{(t)}, \delta\psi_n(Y^*) + \lambda\mathbf{w}^{(t)}\right\rangle \leq 0, \tag{29}$$

where the last line was obtained by using $\mathbf{g}^{(t)} = \delta\psi_n(Y^*) + \lambda\mathbf{w}^{(t)}$ for the exact subgradient case as derived in the paper.

We then get an upper bound on $\eta^{(t)}$:

$$\eta^{(t)} \leq \frac{\left\langle\delta\psi_n(Y^*) + \lambda\mathbf{w}^{(t)}, \delta\psi_n(Y^*) + \lambda\mathbf{w}^{(t)}\right\rangle}{\left(\frac{\lambda}{2} + \rho^{(t)}\right)\left|\left|\delta\psi_n(Y^*) + \lambda\mathbf{w}^{(t)}\right|\right|^2} = \frac{1}{\frac{\lambda}{2} + \rho^{(t)}}. \tag{30}$$

Using the same reasoning for the second case, we can show that $T(\eta^{(t)}) \geq 0$ implies $\eta^{(t)} \geq \frac{1}{\frac{\lambda}{2}+\rho^{(t)}} > 0$, but because we assumed $\eta^{(t)} \leq 0$, this case is not feasible.

We therefore conclude that the solution to Eq. 27 is $\eta^{(t)} \in \left]0; \frac{1}{\frac{\lambda}{2}+\rho^{(t)}}\right]$.

2) Proof of convergence: As shown in Section. B1, $\eta^{(t)}$ close to 0 is a valid solution to Eq. 22, but this could cause the proposed approach to stop prematurely. In order to prevent this issue, we enforce a lower bound on $\eta^{(t)}$ as done in [2]. We thus restrict the bound to $\eta^{(t)} \in \left[\frac{\epsilon}{t}; \frac{1}{\frac{\lambda}{2}+\rho^{(t)}}\right]$ where ϵ is taken to be a very small value so that $\frac{\epsilon}{t} \leq \frac{1}{\frac{\lambda}{2}+\rho^{(t)}}$.
We now show that the step size $\eta^{(t)}$ satisfies the following conditions of convergence:

$$\lim_{T\to+\infty}\sum_{t=1}^{T}\eta^{(t)} = \infty \quad \text{and} \lim_{T\to+\infty} \quad \sum_{t=1}^{T}(\eta^{(t)})^2 < \infty. \tag{31}$$

a) Upper bound
: We start with the upper bound $\frac{1}{\frac{\lambda}{2}+\rho^{(t)}}$. Using the integral test for convergence, we get:

$$\int_{t=1}^{T} \frac{1}{\frac{\lambda}{2}+\rho^{(t)}} dt \;\leq\; \sum_{t=1}^{T} \frac{1}{\frac{\lambda}{2}+\rho^{(t)}}. \tag{32}$$

We also know, recalling that $\rho^{(t)} = Ct$, that

$$\int_{t=1}^{T} \frac{1}{\frac{\lambda}{2}+\rho^{(t)}} dt \;=\; \left.\frac{\log\left(Ct+\frac{\lambda}{2}\right)}{C}\right|_{t=1}^{T} + cst \;\xrightarrow{T\to\infty}\; \infty \tag{33}$$

Therefore

$$\lim_{T\to\infty} \sum_{t=1}^{T} \frac{1}{\frac{\lambda}{2}+\rho^{(t)}} \to \infty, \tag{34}$$

and we see that the upper bound also meets the conditions of convergence.

The same test of convergence can be used to bound the sum of the squares:

$$\int_{t=1}^{T} \frac{1}{\left(\frac{\lambda}{2}+\rho^{(t)}\right)^2} dt \;\leq\; \sum_{t=1}^{T} \frac{1}{\left(\frac{\lambda}{2}+\rho^{(t)}\right)^2}. \tag{35}$$

We also know:

$$\int_{t=1}^{T} \frac{1}{\left(\frac{\lambda}{2}+\rho^{(t)}\right)^2} dt = \left.-\frac{1}{C(Ct+\frac{\lambda}{2})}\right|_{t=1}^{T} + cst. \tag{36}$$

Assuming $\lambda > 0$, we therefore get

$$\lim_{T\to+\infty} \sum_{t=1}^{T} \frac{1}{\left(\frac{\lambda}{2}+\rho^{(t)}\right)^2} < \infty \tag{37}$$

b) Lower bound
: Using the integral test for the lower bound, we also get:

$$\lim_{T\to+\infty} \sum_{t=1}^{T} \frac{\epsilon}{t} = \infty \quad \text{and} \quad \lim_{T\to+\infty} \sum_{t=1}^{T} \left(\frac{\epsilon}{t}\right)^2 < \infty, \tag{38}$$

and we see that the lower bound also meets the conditions of convergence.

We have shown that the lower and upper bounds both satisfy the conditions of convergence from Eq. 31. Therefore any step size $\eta^{(t)}$ within the interval $\eta^{(t)} \in \left[\frac{\epsilon}{t}; \frac{1}{\frac{\lambda}{2}+\rho^{(t)}}\right]$ also satisfies these conditions, thus prooving that Algorithm 2 converges.

REFERENCES

[1] A. Rakhlin, O. Shamir, and K. Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. Technical report, ArXiv, 2012.
[2] H. Zhang, J. Wang, P. Tan, J. Wang, and L. Quan. Learning CRFs for Image Parsing with Adaptive Subgradient Descent. In *ICCV*, 2013.