Constrained deep networks: Lagrangian optimization via Log-barrier extensions

Hoel KERVADEC
Erasmus MC
Netherlands
hoel@kervadec.science

Jose DOLZ
ÉTS Montréal
Canada

Jing YUAN
Xidian University
China

Christian DESROSIERS
ÉTS Montréal
Canada

Eric GRANGER
ÉTS Montréal
Canada

Ismail BEN AYED
ÉTS Montréal
Canada

***Abstract*—This study investigates imposing hard inequality constraints on the outputs of convolutional neural networks (CNN) during training. Several recent works showed that the theoretical and practical advantages of Lagrangian optimization over simple penalties do not materialize in practice when dealing with modern CNNs involving millions of parameters. Therefore, constrained CNNs are typically handled with penalties. We propose *log-barrier extensions*, which approximate Lagrangian optimization of constrained-CNN problems with a sequence of unconstrained losses. Unlike standard interior-point and log-barrier methods, our formulation does not need an initial feasible solution. The proposed extension yields an upper bound on the duality gap—generalizing the result of standard log-barriers—and yielding sub-optimality certificates for feasible solutions. While sub-optimality is not guaranteed for non-convex problems, this result shows that log-barrier extensions are a principled way to approximate Lagrangian optimization for constrained CNNs via *implicit* dual variables. We report weakly supervised image segmentation experiments, with various constraints, showing that our formulation outperforms substantially the existing constrained-CNN methods, in terms of accuracy, constraint satisfaction and training stability, more so when dealing with a large number of constraints.**

***Index Terms*—Constrained CNNs, image segmentation**

I. INTRODUCTION

Imposing prior knowledge in the form of hard constraints on the output of deep convolutional neural networks (CNNs) is useful in a breadth of learning and vision problems. For example, in semi- and weakly-supervised learning, structured prediction or multi-task learning, a set of natural prior-knowledge constraints is available. Such additional knowledge may come from domain experts, for example. In semi-supervised image segmentation, for instance, several recent works [7], [12], [18] showed that imposing domain-specific knowledge on the network's predictions at unlableled data points acts as a powerful regularizer, boosting significantly the performances when the amount of labeled data is limited. Specifically, the authors of [7], [18] added priors on the sizes of the target regions, achieving good performances with only fractions of labels. Such constraints are highly relevant in medical imaging [10], and can mitigate the lack of full annotations[1]. Similar experimental observations were made in other application areas of semi-supervised learning. For example, in natural language processing, the authors of [12], showed that embedding prior-knowledge constraints on unlabled data can yield significant boosts in performances. 3D human pose estimation from a single view [11] is another application where task-specific prior constraints arise naturally, e.g., symmetry constraints encoding that the two arms should have the same length.

As pointed out in several studies [7], [11], [12], [14], [15], [18], imposing hard constraints on modern deep CNNs involving millions of parameters is challenging, even when the constraints are convex with respect to the outputs of the network. In modern deep networks, constraints are commonly handled with *penalties* for their simplicity, and despite their well-known limitations. Standard Lagrangian-dual optimization has been largely avoided and, as discussed in [11], [14], [15], this might be explained by the computational complexity and stability/convergence issues caused by alternating between stochastic optimization and *explicit* dual updates/projections.

Interior-point and log-barrier methods can approximate Lagrangian optimization by starting from a feasible solution and solving unconstrained problems, while completely avoiding explicit dual steps and projections. Unfortunately, despite their well-established advantages over penalties, such standard log-barriers were not used before in deep CNNs because finding a feasible set of initial network parameters is not trivial, and is itself a challenging constrained-CNN problem.

We propose *log-barrier extensions*, which approximate Lagrangian optimization of constrained-CNN problems with a sequence of unconstrained losses, removing the need for an initial feasible set of network parameters. The extensions yield a duality-gap bound, which generalizes the standard duality-gap result of log-barriers, yielding sub-optimality certificates for feasible solutions in the case of convex losses. While sub-optimality is not guaranteed for non-convex problems, this result shows that log-barrier extensions are a principled way to approximate Lagrangian optimization for constrained CNNs via *implicit* dual variables. This addresses the well-known limitations of penalty methods and, at the same time, removes the explicit dual updates of Lagrangian optimization. We report comprehensive weakly supervised segmentation experiments, with various constraints, showing that our formulation outperforms the existing constrained-CNN methods, in terms of accuracy, constraint satisfaction and training stability, more

[1] In semantic segmentation, full supervision involves annotating all pixels in each training image, a problem amplified when annotations require expert knowledge or involves volumetric data as in medical imaging.

ISBN: 978-1-6654-6798-8

so when dealing with a large number of constraints.

II. Method

A. Preliminaries

Let $\mathcal{D} = \{I^1, ..., I^N\}$ denotes a partially labeled set of N training images, and $S_{\boldsymbol{\theta}} = \{s^1_{\boldsymbol{\theta}}, ..., s^N_{\boldsymbol{\theta}}\}$ denotes the associated predicted networks outputs in the form of softmax probabilities, for both unlabeled and labeled data points, with $\boldsymbol{\theta}$ the neural-network weights. These could be class probabilities or dense pixel-wise probabilities in the case of semantic image segmentation. We address constrained problems of the following general form:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \mathcal{E}(\boldsymbol{\theta}) \\ s.t. \quad & f_1(s^n_{\boldsymbol{\theta}}) \leq 0, \quad n = 1, \ldots, N, \\ & \ldots \\ & f_P(s^n_{\boldsymbol{\theta}}) \leq 0, \quad n = 1, \ldots, N. \end{aligned} \tag{1}$$

where $\mathcal{E}(\boldsymbol{\theta})$ is some standard loss over the set of labeled data points—e.g., cross-entropy—and each f_i is some differentiable function, which we want to constrain for each data point n. Inequality constraints of the general form in Eq. (1) can embed very useful prior knowledge on the network's predictions for unlabeled pixels. Assume, for instance, in the case of image segmentation, that we have prior knowledge about the size of the target region (i.e., class) k. Such a knowledge can be in the form of lower or upper bounds on region size, which is common in medical image segmentation problems [2], [7], [13]. In this case, $I^n : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ could be a partially labeled or unlabeled image, with Ω the spatial support of the image, and $s^n_{\boldsymbol{\theta}} \in [0,1]^{K \times |\Omega|}$ its predicted mask. This matrix contains the softmax probabilities for each pixel $p \in \Omega$ and each class k, which we denotes $s^n_{k,p,\boldsymbol{\theta}}$. A constraint in the form of $f_i(s^n_{\boldsymbol{\theta}}) = \sum_{p \in \Omega} s^n_{k,p,\boldsymbol{\theta}} - a \leq 0$ enforces an upper limit a on the size of target region k.

B. Log-barrier extensions

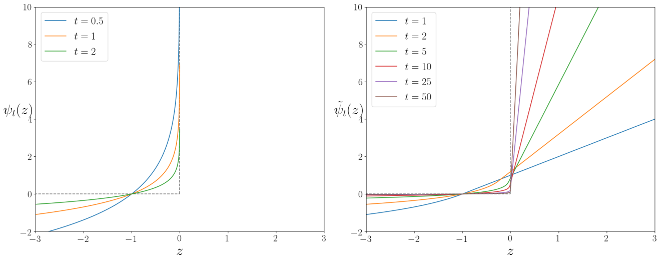
We propose the following unconstrained loss for approximating Lagrangian optimization of constrained problem (1):

$$\min_{\boldsymbol{\theta}} \quad \mathcal{E}(\boldsymbol{\theta}) + \sum_{i=1}^{P} \sum_{n=1}^{N} \tilde{\psi}_t \left(f_i(s^n_{\boldsymbol{\theta}}) \right), \tag{2}$$

where $\tilde{\psi}_t$ is our *log-barrier extension*, which is convex, continuous and twice-differentiable:

$$\tilde{\psi}_t(z) = \begin{cases} -\frac{1}{t} \log(-z) & \text{if } z \leq -\frac{1}{t^2}, \\ tz - \frac{1}{t} \log(\frac{1}{t^2}) + \frac{1}{t} & \text{otherwise.} \end{cases} \tag{3}$$

The standard log-barrier ψ_t and our proposed extension $\tilde{\psi}_t$ are depicted in sub-Figures 1a and 1b respectively. Similarly to the standard log-barrier, when $t \to +\infty$, our extension (3) can be viewed a smooth approximation of hard indicator function H. However, a very important difference is that the domain of our extension $\tilde{\psi}_t$ is not restricted to feasible points $\boldsymbol{\theta}$. Therefore, our approximation (2) removes completely the requirement for explicit Lagrangian-dual optimization for finding a feasible set of network parameters. In our case, the inequality constraints are fully handled within stochastic optimization, as in standard unconstrained losses, avoiding completely gradient ascent iterates and projections over *explicit* dual variables.



(a) Standard log-barrier (b) Log-barrier *extension*

Fig. 1: Plots of the functions, for varying t.

In our approximation in Eq. (2), the Lagrangian dual variables for the initial inequality-constrained problem of (1) are *implicit*. We show the following duality-gap bound, which yields sub-optimality certificates for feasible solutions of our approximation in (2). This result[2] can be viewed as an extension of the standard result in [1][page 566], which expresses the duality-gap as a function of t for the log-barrier function.

Proposition 1. *Let $\boldsymbol{\theta}^*$ be the solution of problem* (2) *and $\boldsymbol{\lambda}^* \in \mathbb{R}^{P \times N}$ the corresponding vector of implicit Lagrangian dual variables given by:*

$$\lambda^*_{i,n} = \begin{cases} -\frac{1}{t f_i(s^n_{\boldsymbol{\theta}^*})} & \textit{if } f_i(s^n_{\boldsymbol{\theta}^*}) \leq -\frac{1}{t^2}, \\ t & \textit{otherwise.} \end{cases} \tag{4}$$

Then, we have the following upper bound on the duality gap associated with primal $\boldsymbol{\theta}^*$ and implicit dual feasible $\boldsymbol{\lambda}^*$ for the initial inequality-constrained problem (1)*:*

$$\mathcal{E}(\boldsymbol{\theta}^*) - g(\boldsymbol{\lambda}^*) \leq PN/t.$$

Proof:

Let $\boldsymbol{\theta}^*$ be the solution of problem (2) and $\boldsymbol{\lambda}^* \in \mathbb{R}^{P \times N}$ the corresponding vector of implicit dual variables given by (4).

We assume that $\boldsymbol{\theta}^*$ verifies approximately[3] the optimality condition for a minimum of (2):

$$\nabla \mathcal{E}(\boldsymbol{\theta}^*) + \sum_{i=1}^{P} \sum_{n=1}^{N} \tilde{\psi'}_t \left(f_i(s^n_{\boldsymbol{\theta}^*}) \right) \nabla f_i(s^n_{\boldsymbol{\theta}^*}) \approx 0 \tag{5}$$

It is easy to verify that each dual variable $\lambda^*_{i,n}$ corresponds to the derivative of the log-barrier extension at $f_i(S_{\boldsymbol{\theta}^*})$:

$$\lambda^*_{i,n} = \tilde{\psi'}_t \left(f_i(s^n_{\boldsymbol{\theta}^*}) \right)$$

[2] The result applies to the general context of convex optimization. In deep CNNs, of course, a feasible solution of our approximation may not be unique and is not guaranteed to be a global optimum as $\mathcal{E}$ and the constraints are not convex.

[3] When optimizing unconstrained loss via stochastic gradient descent (SGD), there is no guarantee that the obtained solution verifies exactly the optimality conditions.

Therefore, Eq. (5) means that $\boldsymbol{\theta}^*$ verifies approximately the optimality condition for the Lagrangian corresponding to the original inequality-constrained problem in Eq. (1) when $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$:

$$\nabla \mathcal{E}(\boldsymbol{\theta}^*) + \sum_{i=1}^{P} \sum_{n=1}^{N} \lambda_{i,n}^* \nabla f_i(s_{\boldsymbol{\theta}^*}^n) \approx 0 \qquad (6)$$

It is also easy to check that the implicit dual variables defined in (4) corresponds to a feasible dual, i.e., $\boldsymbol{\lambda}^* > 0$ element-wise. Therefore, the dual function evaluated at $\boldsymbol{\lambda}^* > 0$ is:

$$g(\boldsymbol{\lambda}^*) = \mathcal{E}(\boldsymbol{\theta}^*) + \sum_{i=1}^{P} \sum_{n=1}^{N} \lambda_{i,n}^* f_i(s_{\boldsymbol{\theta}^*}^n),$$

which yields the duality gap associated with primal-dual pair $(\boldsymbol{\theta}^*, \boldsymbol{\lambda}^*)$:

$$\mathcal{E}(\boldsymbol{\theta}^*) - g(\boldsymbol{\lambda}^*) = -\sum_{i=1}^{P} \sum_{n=1}^{N} \lambda_i^* f_i(s_{\boldsymbol{\theta}^*}^n). \qquad (7)$$

Now, to prove that this duality gap is upper-bounded by PN/t, we consider three cases for each term in the sum in (7) and verify that, for all the cases, we have $\lambda_{i,n}^* f_i(s_{\boldsymbol{\theta}^*}^n) \geq -\frac{1}{t}$.

- $f_i(s_{\boldsymbol{\theta}^*}^n) \leq -\frac{1}{t^2}$: In this case, we can verify that $\lambda_{i,n}^* f_i(s_{\boldsymbol{\theta}^*}^n) = -\frac{1}{t}$ using the first line of (4).
- $-\frac{1}{t^2} \leq f_i(s_{\boldsymbol{\theta}^*}^n) \leq 0$: In this case, we have $\lambda_{i,n}^* f_i(s_{\boldsymbol{\theta}^*}^n) = t f_i(s_{\boldsymbol{\theta}^*}^n)$ from the second line of (4). As t is strictly positive and $f_i(s_{\boldsymbol{\theta}^*}^n) \geq -\frac{1}{t^2}$, we have $t f_i(s_{\boldsymbol{\theta}^*}^n) \geq -\frac{1}{t}$, which means $\lambda_{i,n}^* f_i(s_{\boldsymbol{\theta}^*}^n) \geq -\frac{1}{t}$.
- $f_i(s_{\boldsymbol{\theta}^*}^n) \geq 0$: In this case, $\lambda_{i,n}^* f_i(s_{\boldsymbol{\theta}^*}^n) = t f_i(s_{\boldsymbol{\theta}^*}^n) \geq 0 > -\frac{1}{t}$ because t is strictly positive.

In all the three cases, we have $\lambda_{i,n}^* f_i(s_{\boldsymbol{\theta}^*}^n) \geq -\frac{1}{t}$. Summing this inequality over i gives:

$$-\sum_{i=1}^{P} \sum_{n=1}^{N} \lambda_{i,n}^* f_i(s_{\boldsymbol{\theta}^*}^n) \leq \frac{PN}{t}.$$

Using this inequality in (7) yields the following upper bound on the duality gap associated with primal $\boldsymbol{\theta}^*$ and implicit dual feasible $\boldsymbol{\lambda}^*$ for the original inequality-constrained problem:

$$\mathcal{E}(\boldsymbol{\theta}^*) - g(\boldsymbol{\lambda}^*) \leq PN/t.$$

□

This bound yields sub-optimality certificates for feasible solutions of our approximation in (2). If the solution $\boldsymbol{\theta}^*$ that we obtain from our unconstrained problem (2) is feasible, i.e., it satisfies constraints $f_i(s_{\boldsymbol{\theta}^*}^n) \leq 0$, $\forall i, \forall n$, then $\boldsymbol{\theta}^*$ is PN/t-suboptimal for the original inequality constrained problem: $\mathcal{E}(\boldsymbol{\theta}^*) - \mathcal{E}^* \leq PN/t$. In deep CNNs, of course, a feasible solution for our approximation may not be unique and is not guaranteed to be a global optimum as $\mathcal{E}$ and the constraints are not convex.

From Proposition 1, the following important fact follows immediately: If the solution $\boldsymbol{\theta}^*$ that we obtain from unconstrained problem (2) is feasible and global, then it is PN/t-suboptimal for constrained problem (1): $\mathcal{E}(\boldsymbol{\theta}^*) - \mathcal{E}^* \leq PN/t$.

Similarly to the standard log-barrier algorithm, we use a varying parameter t. At training time, we optimize a sequence of losses of the form (2) and increase gradually the value t by a factor μ. The network parameters obtained for the current t and epoch are used as a starting point for the next t and epoch. This effectively "*raises*" the barrier over time. We can summarize the fundamental differences between our log-barrier extension and a standard penalty function as follows:

A penalty does not act as a barrier near the boundary of the feasible set, i.e., a satisfied constraint yields null penalty and gradient. Therefore, at a given gradient update, there is nothing that prevents a satisfied constraint from being violated, causing oscillations between competing constraints and making the training unstable. On the contrary, the strictly positive gradient of our log-barrier extension gets higher when a satisfied constraint approaches violation during optimization, pushing it back towards the feasible set.

Another fundamental difference is that the derivatives of our log-barrier extensions yield the implicit dual variables in Eq. (4), with sub-optimality and duality-gap guarantees, which is not the case for penalties. Therefore, our log-barrier extension mimics Lagrangian optimization, but with implicit rather than explicit dual variables.

III. Experiments

Most of the existing methods—and the proposed log-barrier—are compatible with any differentiable function f_i, including non-linear and fractional terms, as in Eqs. (8) and (9) introduced further in the paper. However, we hypothesize that our log-barrier extension is better for handling the interplay between multiple competing constraints. To validate this hypothesis, we compare all strategies on the image segmentation tasks with constraints related to region size and location. As baselines we compare to a direct Lagrangian method, and a recent modification of the Lagrangian [12].

a) Region-size constraint: We define the size (or volume) of a segmentation for class k as the sum of its softmax predictions over the image domain:

$$\mathcal{V}_{k,\boldsymbol{\theta}}^n = \sum_{p \in \Omega} s_{k,p,\boldsymbol{\theta}}^n. \qquad (8)$$

We use the following inequality constraints on region size: $0.9\tau_{\mathcal{V}_k^n} \leq \mathcal{V}_{k,\boldsymbol{\theta}}^n \leq 1.1\tau_{\mathcal{V}_k^n}$, where, similarly to the experiments in [7], $\tau_{\mathcal{V}_k^n} = \sum_{p \in \Omega} y_{k,p}^n$ is determined from the ground truth y^n of each image.

b) Region-centroid constraints: The centroid of the predicted region can be computed as a weighted average of the pixel coordinates:

$$\mathcal{C}_{k,\boldsymbol{\theta}}^n = \frac{\sum_{p \in \Omega} s_{k,p,\boldsymbol{\theta}}^n c_p}{\sum_{p \in \Omega} s_{k,p,\boldsymbol{\theta}}^n}, \qquad (9)$$

where $c_p \in \mathbb{N}^2$ are the pixel coordinates on a 2D grid. We constrain the position of the centroid in a box around the ground-truth centroid: $\tau_{\mathcal{C}_k^n} - 20 \leq \mathcal{C}_{k,\boldsymbol{\theta}}^n \leq \tau_{\mathcal{C}_k^n} + 20$, with $\tau_{\mathcal{C}_k^n} = \frac{\sum_{p \in \Omega} y_{k,p}^n c_p}{\sum_{p \in \Omega} y_{k,p}^n}$ corresponding to the bound values associated with each image.

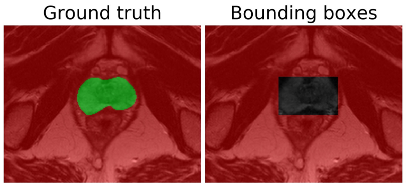


Fig. 2: Full mask of the prostate (*left*) and the box annotations (*right*). The background is in red and the foreground in green. No color means that no information is provided.

c) Bounding box tightness prior: This prior [5], [8] assumes that any horizontal or vertical line inside the bounding box of an object of class k will eventually cross the object. This can be generalized with segments of width w inside the box, that will cross at least w times the object. This prior can be easily reformulated as constraints. If $\mathcal{S}_L^n := \{s_l^n\}$ denotes the set of parallel segments to the sides of the bounding box for sample n, the following set of inequality constraints is trivial to define:

$$\sum_{p\in s_l^n} y_{k,p}^n \geq w \qquad \forall s_l^n \in \mathcal{S}_L^n, \forall n \in \mathcal{D}. \tag{10}$$

If we define the inside of the bounding box as Ω_F, and the outside as Ω_B (such as $\Omega = \Omega_F \cup \Omega_B$ and $\Omega_F \cap \Omega_B = \{\emptyset\}$), we can define two other useful constraints for each image:

$$\sum_{p\in\Omega_B} s_{k,p,\boldsymbol{\theta}}^n \leq 0 \qquad \forall n \in \mathcal{D}, \tag{11}$$

$$\sum_{p\in\Omega} s_{k,p,\boldsymbol{\theta}}^n \leq |\Omega_F| \qquad \forall n \in \mathcal{D}. \tag{12}$$

This setting is a good benchmark to evaluate the interplay of numerous, competing constraints simultaneously.

A. Datasets and evaluation metrics

Experiments were performed on two different segmentation scenarios using synthetic and medical images.

Synthetic images: We randomly generated a synthetic dataset composed of 1100 images with two different circles of the same size but different intensity values, where the darker circle is the target region (Fig. 3, first column). Furthermore, different levels of Gaussian noise were added to the images. We employed 1000 images for training and 100 for validation. We test the combinations of contraints (8) and (9).

Medical images: We use the dataset from the MICCAI 2012 prostate segmentation challenge [9], PROMISE12. This dataset contains Magnetic Resonance (MR) images from 50 patients, from which we employ 10 patients for validation and use the rest for training. We test the combinations of constraints (10), (11) and (12), with bounding boxes derived from the ground truth (illustrated in Figure 2).

Evaluation: We resort to the common Dice index (DSC) = $\frac{2|S\bigcap Y|}{|S|+|Y|}$ to evaluate predicted segmentations. Furthermore, we evaluate the effectiveness and stability of the constrained optimization methods. To this end, we first compute at each epoch the percentage of constraints that are satisfied. Second, we measure the stability of the constraints, i.e., the percentage of constraints satisfied at epoch t that are still satisfied at epoch $t+1$. And last, we measure the time needed to train a single epoch, including the dual update for the Standard Lagrangian and ReLU Lagrangian [12].

TABLE I: Mean DSC and standard deviation of the last 10 epochs on the validation on the toy example and PROMISE12 datasets.

Method	Synthetic dataset	PROMISE12
Standard Lagrangian	0.005 (0.014)	0.000 (0.000)
ReLU Lagrangian [12]	0.798 (0.006)	0.000 (0.000)
Penalty [4], [7]	0.712 (0.022)	0.000 (0.000)
Log-barrier extensions (ours)	**0.945** (0.001)	**0.813** (0.024)
Full supervision	0.998 (0.000)	0.880 (0.001)

The code is publicly available[4], and contains all relevant implementation details, hyper-parameters, and running scripts for easier reproducibility.

B. Results

a) Quantitative results: Results in terms of DSC are reported in Table I. The first thing we can observe on the synthetic dataset is that the standard Lagrangian, despite the introduction of a dedicated learning rate for its $\boldsymbol{\lambda}$ update, is not able to learn when multiple constraints are in competition, i.e, DSC of 0.005 in the synthetic example. In addition, the ReLU Lagrangian approach proposed by [12] can better handle multiple constraints than a simple penalty [4], [7].

With the high number of constraints and trivial solutions to balance, the proposed log-barrier extension learns successfully based on the information given by the constraints, compared to the other methods, achieving the best DSC across the two settings, and, more importantly, is the only method managing to predict non-empty segmentations on the medical task.

The very poor performance of penalty-based methods can be explained by the high-gradients generated when constraints are not satisfied, which leads to big and simplistic updates.

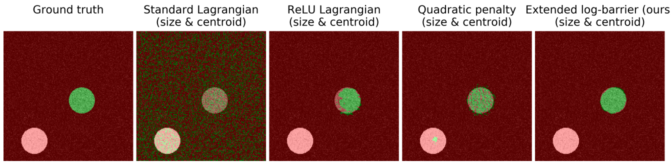


Fig. 3: Results on the synthetic dataset (background in red and foreground in green).

b) Qualitative results: A visual comparison on the synthetic dataset is depicted in Figure 3. In this figure we can first observe that standard Lagrangian generates noisy segmentations, which is in line with the quantitative results reported in Table I. Both ReLU Lagrangian [12] and penalty-based methods obtain better target segmentations. Nevertheless, they

[4] https://github.com/LIVIAETS/extended_logbarrier

cannot handle efficiently the interplay between multiple constraints. Meanwhile, the proposed extended log-barrier demonstrates a strong ability to handle several constraints simultaneously, which is reflected in the circle segmentation close to the ground truth.

c) Constraints satisfaction and stability: The metrics over training epochs are shown in Fig. 4. We can notice that on top of the better absolute performances, the proposed log-barrier extension is also more stable during training, both in performance and constraints satisfaction.

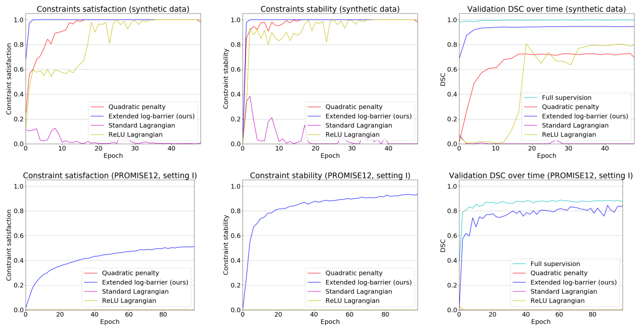


Fig. 4: Constraints satisfaction, stability and DSC evolution on different settings. Best viewer in colors.

d) Computational cost and efficiency: Penalties and the proposed log-barrier extension have negligible cost compared to optimizing the base-loss $\mathcal{E}(\boldsymbol{\theta})$ alone (up to 5% slowdown when the number of constraints becomes very high). In contrast, Lagrangian methods incur in higher computational cost. For example, in the standard and ReLU Lagrangian, it amounts to nearly a 25% slowdown (due to the extra loop over the training set to perform the $\boldsymbol{\lambda}$ update).

IV. Conclusion

We proposed *log-barrier extensions*, which approximate Lagrangian optimization of constrained-CNN problems with a sequence of unconstrained losses. Our formulation relaxes the need for an initial feasible solution, unlike standard interior-point and log-barrier methods. This makes it convenient for deep networks. We also provided an upper bound on the duality gap for our proposed extensions, thereby generalizing the duality-gap result of standard log-barriers and showing that our formulation has dual variables that mimic implicitly (without dual projections/steps) Lagrangian optimization. Therefore, our implicit Lagrangian formulation can be fully handled with SGD, the workhorse of deep networks. We reported constrained-CNN experiments, showing that log-barrier extensions outperform several other types of Lagrangian methods and penalties, in terms of accuracy and training stability. Log-barrier extensions can be useful in breadth of problems in vision and learning, where constraints occur naturally. This include, for instance, adversarial robustness [16], stabilizing the training of GANs [3], domain adaptation for segmentation [17], pose-constrained image generation [6], 3D human pose estimation [11], deep reinforcement learning [4] and natural language processing [12]. To our knowledge, those constraints are typically handled with basic penalties; it will therefore be interesting to investigate log-barrier extensions in these diverse contexts.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] Lena Gorelick, Frank R. Schmidt, and Yuri Boykov. Fast trust region for segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1714–1721, 2013.

[3] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *Neural Information Processing Systems (NIPS)*, pages 5767–5777, 2017.

[4] Frank S. He, Yang Liu, Alexander G. Schwing, and Jian Peng. Learning to play in a day: Faster deep reinforcement learning by optimality tightening. In *International Conference on Learning Representations (ICLR)*, pages 1–13, 2017.

[5] Cheng-Chun Hsu, Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised instance segmentation using the bounding box tightness prior. In *Advances in Neural Information Processing Systems*, pages 6582–6593, 2019.

[6] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, Lianhui Qin, Xiaodan Liang, Haoye Dong, and Eric P. Xing. Deep generative models with learnable knowledge constraints. In *Neural Information Processing Systems (NeurIPS)*, pages 10522–10533, 2018.

[7] Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. Constrained-cnn losses for weakly supervised segmentation. *Medical image analysis*, 54:88–99, 2019.

[8] Victor Lempitsky, Pushmeet Kohli, Carsten Rother, and Toby Sharp. Image segmentation with a bounding box prior. In *2009 IEEE 12th international conference on computer vision*, pages 277–284. IEEE, 2009.

[9] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

[10] Geert J. S. Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[11] Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Imposing Hard Constraints on Deep Networks: Promises and Limitations. In *CVPR Workshop on Negative Results in Computer Vision*, pages 1–9, 2017.

[12] Yatin Nandwani, Abhishek Pathak, Parag Singla, et al. A primal dual formulation for deep learning with constraints. In *Advances in Neural Information Processing Systems*, pages 12157–12168, 2019.

[13] Marc Niethammer and Christopher Zach. Segmentation with area constraints. *Medical Image Analysis*, 17(1):101–112, 2013.

[14] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *International Conference on Computer Vision (ICCV)*, pages 1796–1804, 2015.

[15] Sathya N Ravi, Tuan Dinh, Vishnu Suresh Lokhande, and Vikas Singh. Explicitly imposing constraints in deep networks via conditional gradients gives improved generalization and faster convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4772–4779, 2019.

[16] J. Rony, L. G. Hafemann, L. S. Oliveira, I. Ben Ayed, R. Sabourin, and E. Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1–10, 2019.

[17] Yang Zhang, Philip David, Hassan Foroosh, and Boqing Gong. A curriculum domain adaptation approach to the semantic segmentation of urban scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[18] Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *International Conference on Computer vision (ICCV)*, pages 1–10, 2019.