# Regularized Frank-Wolfe for Dense CRFs: Generalizing Mean Field and Beyond

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

We introduce *regularized Frank-Wolfe*, a general and effective algorithm for inference and learning of dense conditional random fields (CRFs). The algorithm optimizes a nonconvex continuous relaxation of the CRF inference problem using vanilla Frank-Wolfe with approximate updates, which are equivalent to minimizing a regularized energy function. Our proposed method is a generalization of existing algorithms such as mean field or concave-convex procedure. This perspective not only offers a unified analysis of these algorithms, but also allows an easy way of exploring different variants that potentially yield better performance. We illustrate this in our empirical results on standard semantic segmentation datasets, where several instantiations of our regularized Frank-Wolfe outperform mean field inference, both as a standalone component and as an end-to-end trainable layer in a neural network. We also show that dense CRFs, coupled with our new algorithms, produce significant improvements over strong CNN baselines.

## 1 Introduction

Fully-connected or *dense* conditional random fields (CRFs) [33]—combined with strong pixel-level classifiers such as a convolutional neural network (CNN) [24, 40]—have been a highly-successful paradigm for semantic segmentation. Top-performing systems on the PASCAL VOC benchmark [21] used to include a CRF as either a post-processing step [12, 13, 14, 19, 41, 42, 43] or a trainable component [4, 47, 58, 61, 63, 66]. However, as CNNs got stronger, the improvements that CRFs brought decreased over time, and as a result they fell out of favor since 2017 [43].

In this paper, we revisit dense CRFs with two contributions. First, on the theoretical side, we propose *regularized Frank-Wolfe*, a new class of algorithms for inference and learning of CRFs that perform better than the popular *mean field* [33, 34, 53]—the method of choice in the aforementioned works. Regularized Frank-Wolfe optimizes a nonconvex continuous relaxation of the CRF inference problem (§2) by performing approximate conditional-gradient updates (§3.1), which is equivalent to minimizing a regularized energy using the generalized Frank-Wolfe method [50] (§3.2). Several of its instantiations lead to new algorithms that have not been studied before in the MAP inference literature (§3.3). Moreover, we show that it also includes several existing methods, including mean field and the concave-convex procedure [65], as special cases (§3.4). This generalized perspective allows a unified analysis of all these old and new algorithms in a single framework (§4). In particular, we show that they achieve a sublinear rate of convergence $\mathcal{O}(1/\sqrt{k})$ for suitable stepsize schemes, and in certain cases (such as strongly-convex regularizer or concave energy) this can be improved to $\mathcal{O}(1/k)$ (§4.1). Furthermore, we provide a tightness analysis for the resulting nonconvex relaxation of the regularized energy, recovering some existing tightness results [9, 39, 55] as special cases (§4.2). The proposed algorithms are easy to implement, converge quickly in practice, and have (sub)differentiable iterates. Such properties are important for successful *gradient-based learning* via backpropagation [45, 56].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Our second contribution lies on the practical side. In addition to mean field and regularized Frank-Wolfe variants, we re-implement several existing first-order inference methods [38, 39, 46]—those that are amenable to gradient-based learning—for comparison. Remarkably, we find that dense CRFs can still achieve important improvements over the strong DeepLabv3+ [16] CNN model for all these solvers (§5). In particular, our best variant of regularized Frank-Wolfe achieves a mean intersection-over-union (mIoU) score of 88.0 on the PASCAL VOC test set (§5.3), improving over DeepLabv3+. We hope that these encouraging results could attract interest from the community in considering dense CRFs (again) for tasks such as semantic segmentation. Our source code will be made publicly available under the GNU general public license for this purpose.

# 2 Background

## 2.1 Inference in CRFs

Let $\mathbf{s} \in \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$ denote an assignment to $n$ discrete random variables $S_1, \ldots, S_n$, where each variable $S_i$ takes values in a finite set of states (or *labels*) $\mathcal{S}_i$. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph of $n$ nodes ($\mathcal{V} = \{1, 2, \ldots, n\}$). A Markov random field (MRF) defined by $\mathcal{G}$ encodes a family of joint distributions that can be factorized as follows, where $\phi_i : \mathcal{S}_i \to \mathbb{R}_+$ and $\phi_{ij} : \mathcal{S}_i \times \mathcal{S}_j \to \mathbb{R}_+$ are the so-called *unary* and *pairwise* (respectively) potential functions, and $Z$ is a normalization factor:

$$p(\mathbf{s}) = \frac{1}{Z} \prod_{i \in \mathcal{V}} \phi_i(s_i) \prod_{ij \in \mathcal{E}} \phi_{ij}(s_i, s_j). \tag{1}$$

Note that (1) can also include conditional distributions, i.e., $p(\mathbf{s}|\mathbf{o})$ with observed variables $\mathbf{o}$. In this case the potentials may also depend on $\mathbf{o}$, e.g., $\phi_i(s_i; \mathbf{o})$, and this model is referred to as conditional random field (CRF) [36]. We will present later (§5.1) such a model for image segmentation. In the following, we use MRF and CRF interchangeably. We assume further that all nodes have the same set of labels: $\mathcal{S}_i = \mathcal{S}\ \forall i$, with cardinality $d = |\mathcal{S}|$. It is convenient to express $p(\mathbf{s})$ as $\frac{1}{Z}\exp(-e(\mathbf{s}))$, where the so-called *energy* $e(\mathbf{s})$ is defined as

$$e(\mathbf{s}) = \sum_{i \in \mathcal{V}} \theta_i(s_i) + \sum_{ij \in \mathcal{E}} \theta_{ij}(s_i, s_j), \quad \text{with } \theta_i(s_i) = -\log \phi_i(s_i) \text{ (idem for } \theta_{ij}). \tag{2}$$

The task of maximum a posteriori (MAP) inference consists in finding the most probable joint assignment, also known as *energy minimization* (which is NP-Hard in general [59]):

$$\mathbf{s}^* = \operatorname*{argmax}_{\mathbf{s} \in \mathcal{S}^n} p(\mathbf{s}) = \operatorname*{argmin}_{\mathbf{s} \in \mathcal{S}^n} e(\mathbf{s}). \tag{3}$$

## 2.2 Continuous relaxation of MAP inference

Let $x_{is}$ be a binary variable such that $x_{is} = 1$ iff label $s$ is assigned to node $i$. Then, $\mathbf{x}_i = (x_{is})_{s \in \mathcal{S}} \in \{0, 1\}^d$ denotes the one-hot vector for node $i$. Let $\mathbf{x} \in \{0, 1\}^{nd}$ be the concatenation of all $\mathbf{x}_i$, and let $\boldsymbol{\theta}_i = (\theta_i(s))_{s \in \mathcal{S}} \in \mathbb{R}^d$, $\boldsymbol{\Theta}_{ij} = (\theta_{ij}(s,t))_{t \in \mathcal{S}}^{s \in \mathcal{S}} \in \mathbb{R}^{d \times d}$. The energy (2) is then

$$E(\mathbf{x}; \boldsymbol{\theta}) = \sum_{i \in \mathcal{V}} \boldsymbol{\theta}_i^\top \mathbf{x}_i + \sum_{ij \in \mathcal{E}} \mathbf{x}_i^\top \boldsymbol{\Theta}_{ij} \mathbf{x}_j, \tag{4}$$

where $\boldsymbol{\theta}$ is a parameter vector composed of all $\boldsymbol{\theta}_i$ and $\boldsymbol{\Theta}_{ij}$. The MAP inference problem (3) transforms to minimizing $E(\mathbf{x}; \boldsymbol{\theta})$ over $\mathbf{x}$. A natural approach is to relax the binary constraint and solve the continuous relaxation $\min_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}; \boldsymbol{\theta})$ with

$$\mathcal{X} = \left\{ \mathbf{x} \in \mathbb{R}^{nd} : \mathbf{x} \geq \mathbf{0}, \mathbf{1}^\top \mathbf{x}_i = 1\ \forall i \in \mathcal{V} \right\}. \tag{5}$$

This continuous relaxation is known to be *tight* [8, 39, 55]. For convenience, we represent the unary potentials as a vector $\mathbf{u}(\boldsymbol{\theta}) = (\boldsymbol{\theta}_i)_{i \in \mathcal{V}} \in \mathbb{R}^{nd}$, and the pairwise potentials as a symmetric $n \times n$ block matrix $\mathbf{P}(\boldsymbol{\theta}) \in \mathbb{R}^{nd \times nd}$, where the $(i, j)$ block is $\boldsymbol{\Theta}_{ij}$. The problem is reduced to

$$\min_{\mathbf{x} \in \mathcal{X}} E(\mathbf{x}; \boldsymbol{\theta}) \triangleq \frac{1}{2} \mathbf{x}^\top \mathbf{P}(\boldsymbol{\theta}) \mathbf{x} + \mathbf{u}(\boldsymbol{\theta})^\top \mathbf{x}. \tag{6}$$

The reason we have made $\boldsymbol{\theta}$ explicit in (4) and (6) is to provide more clarity when discussing the differentiability of their solutions for the *learning* task. Note that an optimal solution $\mathbf{x}^*$ to (4) and (6) is a function of $\boldsymbol{\theta}$, and thus should be written as $\mathbf{x}^*(\boldsymbol{\theta})$. Therefore, when we say a solution is differentiable, it is understood that it is so with respect to $\boldsymbol{\theta}$. When there is no ambiguity, we omit $\boldsymbol{\theta}$ and write simply $E(\mathbf{x})$, $\mathbf{P}$, and $\mathbf{u}$. In the rest of the paper, (6) will be our problem of interest.

### 2.3 Vanilla Frank-Wolfe algorithm for MAP inference

Since the continuous energy is differentiable, it is natural to apply first-order methods such as Frank-Wolfe [22] to solving (6) [39]. Starting from a feasible $\mathbf{x}^0 \in \mathcal{X}$, Frank-Wolfe approximately solves (6) by iterating the following steps, where $\alpha_k \in [0, 1]$ follows some stepsize scheme:

$$\mathbf{p}^k \in \operatorname*{argmin}_{\mathbf{p}\in\mathcal{X}} \left\langle \nabla E(\mathbf{x}^k), \mathbf{p} \right\rangle, \qquad \mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k(\mathbf{p}^k - \mathbf{x}^k). \tag{7}$$

The same idea has also been applied to other types of continuous relaxations of (6), such as linear programming (LP) [49] or convex quadratic programming (QP) [20, 55].

## 3 Regularized Frank-Wolfe for inference

In this section, we introduce the proposed regularized Frank-Wolfe as a general class of algorithms for MAP inference. The motivation and general framework are presented in §3.1 and §3.2. We show that several new inference algorithms can be obtained using different regularizers (§3.3). Moreover, regularized Frank-Wolfe also includes a number of existing algorithms as special cases (§3.4).

### 3.1 A smoothing perspective

We have seen in §2.3 the (vanilla) Frank-Wolfe method for solving MAP inference. Unfortunately, from a *learning* perspective, this algorithm is problematic. Indeed, gradient-based CRF learning is typically done by backpropagating through the optimization steps [34, 57, 66], but the solution $\mathbf{p}^k$ to (7) is piecewise constant, thus its gradient (w.r.t. $\boldsymbol{\theta}$) is zero almost everywhere (Appendix A.1), which makes learning not possible with SGD.[1] This will be illustrated later in the experiments. A potential solution is to add a (typically strongly-)convex regularization term:

$$\mathbf{p}^k \in \operatorname*{argmin}_{\mathbf{p}\in\mathcal{X}} \left\{ \left\langle \nabla E(\mathbf{x}^k), \mathbf{p} \right\rangle + r(\mathbf{p}) \right\}, \qquad \mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k(\mathbf{p}^k - \mathbf{x}^k). \tag{8}$$

With appropriately chosen regularizers, (8) becomes suitable for gradient-based learning.

The technique of approximating an update step by a regularized one is quite standard in the optimization literature. For example, the classical proximal gradient method [46] can be interpreted the same way. Our update (8) is inspired by Nesterov's smoothing [51], and thus is similar in spirit to its many applications [31, 52, 60]. In the next section, we present a different, more general perspective to view (8), which offers more flexibility in designing new algorithms, in making connections to existing ones, and in analyzing their theoretical properties in a unified manner.

### 3.2 A regularized energy perspective

Instead of approximating the local updates of a *given algorithm* (in this case: vanilla Frank-Wolfe) to minimize the *same* objective function, one may choose to keep the algorithm, and approximate instead the objective. At first glance, however, this idea does not seem to be a good one. Indeed, if we replace $E$ by some function $E'$ and apply vanilla Frank-Wolfe, the updates (7) remains piecewise constant, thus we have the same zero-gradient issue. It turns out that, if we choose a more appropriate "given algorithm", this idea can work. Such a choice is the *generalized Frank-Wolfe* algorithm [10, 50].

Consider the following problem, where $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is differentiable but possibly nonconvex, and $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is proper, closed, and convex but possibly non-differentiable:

$$\min_{\mathbf{x}} F(\mathbf{x}) \triangleq f(\mathbf{x}) + g(\mathbf{x}). \tag{9}$$

Generalized Frank-Wolfe solves (9) by iterating

$$\mathbf{p}^k \in \operatorname*{argmin}_{\mathbf{p}} \left\{ \left\langle \nabla f(\mathbf{x}^k), \mathbf{p} \right\rangle + g(\mathbf{p}) \right\}, \qquad \mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k(\mathbf{p}^k - \mathbf{x}^k). \tag{10}$$

If $g$ is the indicator function $\delta_{\mathcal{X}}$ of $\mathcal{X}$ (i.e., $\delta_{\mathcal{X}}(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathcal{X}$ and $\delta_{\mathcal{X}}(\mathbf{x}) = +\infty$ otherwise), then the algorithm clearly reduces to vanilla Frank-Wolfe (7) for $\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$. Therefore, generalized

[1] Blackbox differentiation [54] can deal with this scenario, but it is limited to LP relaxations.

Frank-Wolfe applied to (6) with $f = E$ and $g = \delta_{\mathcal{X}}$ will yield exactly the same updates (7). Now let us apply this algorithm to an approximate objective $E_r(\mathbf{x}) = E(\mathbf{x}) + r(\mathbf{x})$ for some function $r$. Choosing $f = E$ and $g = r + \delta_{\mathcal{X}}$, it is straightforward that (10) reduces to (8). Therefore, we have recovered the same algorithm as in §3.1, but this time through different machinery.

This framework offers a great flexibility as one can choose $f$ and $g$ in many different ways to obtain new algorithms. The only conditions are $f$ being differentiable and $g$ being convex, so that the subproblem in (10) is well-defined and globally solvable.[2] For example, instead of choosing $f = E$ and $g = r + \delta_{\mathcal{X}}$ as above, one can choose $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{P}\mathbf{x}$ and $g(\mathbf{x}) = \mathbf{u}^\top \mathbf{x} + r(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x})$. We will recover later in §3.4 some existing algorithms (as special cases) through this kind of decomposition. Finally, we present Algorithm 1 for (approximately) solving MAP inference (6).

**Algorithm 1** Generic regularized Frank-Wolfe for (approximately) solving MAP inference (6).

1: Choose a regularizer $r$ such that there exist $f$ (differentiable) and $g$ (convex) satisfying $f + g = E + r + \delta_{\mathcal{X}}$. Typically (but not necessarily) $r$ is convex on $\mathcal{X}$ and is constant on $\mathcal{X} \cap \{0,1\}^{nd}$.
2: Initialization: $k \leftarrow 0$, $\mathbf{x}^0 \in \mathcal{X}$, number of iterations $N$.
3: Compute $\mathbf{p}^k \in \operatorname{argmin}_{\mathbf{p}} \left\{ \langle \nabla f(\mathbf{x}^k), \mathbf{p} \rangle + g(\mathbf{p}) \right\}$ and compute the stepsize $\alpha_k$.
4: Update $\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha_k(\mathbf{p}^k - \mathbf{x}^k)$. Let $k \leftarrow k + 1$ and go to Step 3 until $k = N$.
5: Rounding: convert $\mathbf{x}$ to a discrete solution and return.

While the choice of $(r, f, g)$ can be highly flexible, it would make little sense to optimize a function that has nothing to do with the original objective (i.e., the discrete energy). Let $\overline{\mathcal{X}} = \mathcal{X} \cap \{0,1\}^{nd}$ denote the discrete domain of our problem. If we choose $r$ such that it is constant on $\overline{\mathcal{X}}$ (as suggested in Step 1 above), then minimizing $E$ on $\overline{\mathcal{X}}$ is equivalent to minimizing $E + r$ on $\overline{\mathcal{X}}$, and thus Algorithm 1 actually solves the continuous relaxation of a (different) discrete problem that is equivalent to MAP inference. Further discussion on this matter, as well as on the rounding Step 5, are deferred until §4.2.

Finally, we should note that adding a strongly-convex regularizer is not new in the MAP inference literature [29, 31, 49]. In particular, Meshi et al. [49] even applied (vanilla) Frank-Wolfe to optimizing such regularized energy. However, as already explained in the beginning of §3.2, doing that results in zero gradients, thus their algorithm is fundamentally different than ours.

### 3.3 Particular instantiations

The previous section presents regularized Frank-Wolfe as a general algorithm for inference. We now discuss concrete examples of its instantiations. To the best of our knowledge, *all the algorithms presented in this section are new and have not been studied previously in the MAP inference literature*.

**$\ell_2$ Frank-Wolfe** Perhaps the most natural choice is $\ell_2$ regularization. In Algorithm 1, let us choose $f(\mathbf{x}) = E(\mathbf{x})$ and $r(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}\|_2^2$, where $\lambda$ is a positive regularization weight. Let $\Pi_{\mathcal{X}}(\mathbf{v})$ be the projection of a vector $\mathbf{v}$ onto the set $\mathcal{X}$. It can be shown (§A.2) that Step 3 in Algorithm 1 becomes

$$\mathbf{p}^k = \operatorname*{argmin}_{\mathbf{p} \in \mathcal{X}} \left\{ \langle \mathbf{P}\mathbf{x}^k + \mathbf{u}, \mathbf{p} \rangle + \frac{\lambda}{2}\|\mathbf{p}\|_2^2 \right\} = \Pi_{\mathcal{X}}\left( -\frac{1}{\lambda}(\mathbf{P}\mathbf{x}^k + \mathbf{u}) \right) \quad \forall k \geq 0. \tag{11}$$

**Entropy Frank-Wolfe** In Algorithm 1, let us choose $f(\mathbf{x}) = E(\mathbf{x})$ and $r(\mathbf{x}) = -\lambda H(\mathbf{x})$, where $\lambda > 0$ is a regularization weight and $H(\mathbf{x}) = -\sum_{i \in \mathcal{V}} \sum_{s \in \mathcal{S}} x_{is} \log x_{is}$ is the entropy of $\mathbf{x}$ over $\mathcal{X}$. It can be shown (§A.2) that Step 3 in Algorithm 1 becomes

$$\mathbf{p}^k = \operatorname*{argmin}_{\mathbf{p} \in \mathcal{X}} \left\{ \langle \mathbf{P}\mathbf{x}^k + \mathbf{u}, \mathbf{p} \rangle - \lambda H(\mathbf{x}) \right\} = \operatorname{softmax}\left( -\frac{1}{\lambda}(\mathbf{P}\mathbf{x}^k + \mathbf{u}) \right) \quad \forall k \geq 0, \tag{12}$$

where $\mathbf{v} = \operatorname{softmax}(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^{nd}$ means $\mathbf{v} \in \mathbb{R}^{nd}$ and $v_{is} = \frac{\exp(x_{is})}{\sum_{t \in \mathcal{S}} \exp(x_{it})}$ $\forall i \in \mathcal{V}, \forall s \in \mathcal{S}$. The resulting algorithm has a tight connection with (parallel) mean field [33, 34] (§3.4).

**Other variants** One can consider more sophisticated regularizers, e.g., a weighted combination of $\ell_2$ and entropy. Other options include the many different regularizers that have been used in diverse machine learning applications, such as $\ell_p$ norm [52], lasso variants [52], or binary entropy [3]. Although these variants also lead to new MAP inference algorithms, their implementations are more sophisticated since their subproblems (10) require numerical solutions as no closed form ones exist.

[2] Further mild conditions are required for convergence (§4.1).

### 3.4 Recovering existing algorithms as special cases

In addition to the above new algorithms, regularized Frank-Wolfe also includes several existing ones as special cases. We present some of them below and refer to Appendix A.3 for further details.

**Mean field** This is a special case of the above Entropy Frank-Wolfe. Indeed, if we choose $\lambda = 1$ in (12) and a constant stepsize $\alpha_k = 1 \; \forall k \geq 0$ in Algorithm 1, then it is straightforward that this algorithm is reduced to the following update step, where $\mathcal{N}_i$ is the set of neighbors of node $i$:

$$\mathbf{x}^{k+1} = \operatorname{softmax}(-\mathbf{P}\mathbf{x}^k - \mathbf{u}) \iff x_{is}^{k+1} = \tfrac{1}{Z_i} \exp\left( -\theta_i(s) - \textstyle\sum_{j \in \mathcal{N}_i} \sum_{t \in \mathcal{S}} \theta_{ij}(s,t) x_{jt}^k \right) \forall i \in \mathcal{V}, s \in \mathcal{S}.$$

This is precisely a (parallel) mean field update [33, 34]. To conclude, parallel mean field is an instance of Entropy Frank-Wolfe with unit regularization weight and unit stepsize. Interestingly, the update (12) is the well-known softmax function with *temperature* in the deep learning literature [28]. One could have easily come up with such a simple extension of mean field by adding a temperature to softmax (yet surprisingly this has not been tried before), but here we have provided a principled way to achieve that. As shown later in the experiments, with suitable $\lambda$, this extension yields much better results than vanilla mean field. Finally, we should note that the tight connection between mean field and first-order methods has been noticed before. Krähenbühl and Koltun [34] proposed several mean-field-type variants based on the concave-convex procedure [65], while closely similar variants can also be obtained through proximal gradient [2, 6], but unlike our generalized algorithm, these algorithms cannot recover *exactly* the original mean field of Krähenbühl and Koltun [33].

**Concave-convex procedure** CCCP [65] solves (9), assuming $f$ is concave and $g$ is convex, by updating $\mathbf{x}^{k+1}$ as a solution to $-\nabla f(\mathbf{x}^k) \in \partial g(\mathbf{x}^{k+1})$,[3] which is precisely (10) with stepsize $\alpha_k = 1$. We conclude that CCCP is a special case of generalized Frank-Wolfe with $f$ concave and unit stepsize. As a result, many existing CCCP-based inference algorithms [20, 34] can be seen as special cases of regularized Frank-Wolfe. For example, the ones presented by Desmaison et al. [20] are instantiations of the proposed algorithm with either $f(\mathbf{x}) = -\mathbf{x}^\top \operatorname{diag}(\mathbf{c})\mathbf{x}$ and $r(\mathbf{x}) = E(\mathbf{x}) + \mathbf{x}^\top \operatorname{diag}(\mathbf{c})\mathbf{x}$ (where $\mathbf{c} \in \mathbb{R}^{nd}$ is large enough so that $r(\mathbf{x})$ is convex), or $f(\mathbf{x}) = \mathbf{x}^\top(\mathbf{P} - \mathbf{C})\mathbf{x}$ and $r(\mathbf{x}) = \mathbf{u}^\top\mathbf{x} - \mathbf{x}^\top\mathbf{C}\mathbf{x}$ (where $\mathbf{C}$ is some matrix such that $f$ is concave and $r$ is convex). Note that in these instantiations, Step 3 in Algorithm 1 requires a numerical solution (i.e., Algorithm 1 has an inner loop). Finally, all the algorithms presented by Krähenbühl and Koltun [34] are also instantiations of the proposed method because they are based on CCCP. We refer to Appendix A.3 for further details.

**Vanilla Frank-Wolfe** This is trivially a special case of regularized Frank-Wolfe and we briefly discuss it for completeness. Choosing $f(\mathbf{x}) = E(\mathbf{x})$ and $r(\mathbf{x}) = 0$ we obtain the algorithm by Lê-Huu and Paragios [39]. Likewise, the one by Desmaison et al. [20] corresponds to $f(\mathbf{x}) = E(\mathbf{x}) - \mathbf{c}^\top\mathbf{x} + \mathbf{x}^\top \operatorname{diag}(\mathbf{c})\mathbf{x}$ and $r(\mathbf{x}) = 0$, where $\mathbf{c} \in \mathbb{R}^{nd}$ is large enough for $f$ to be convex.

## 4 Theoretical analysis

### 4.1 Convergence

We provide a convergence analysis for the generalized Frank-Wolfe algorithm, and the results for CRF inference special cases will then follow as a consequence. Convergence of vanilla Frank-Wolfe has been well studied in the literature [23, 30, 35, 37]. For generalized Frank-Wolfe, different analyses exist for the case where both $f$ and $g$ in (9) are convex [5, 25, 48, 64]. We are particularly interested in the general case where $f$ **is nonconvex**,[4] as the CRF energy is often highly so in practice. Mine and Fukushima [50] (and subsequently Bredies et al. [11]) proved the global convergence of the algorithm under mild conditions, though no rate of convergence was given. Recently, Beck [7] obtained a rate of convergence $\mathcal{O}(1/\sqrt{k})$ for convex $g$ under adaptive or line-search stepsizes. We extend their analysis with several contributions. First, we include the case where $g$ is strongly convex, which is important as our main variants for inference (e.g., mean field or $\ell_2$-Frank-Wolfe) use strongly-convex regularizers. Second, to also include CCCP [65] as a special case, we relax their Lipschitz smoothness assumption on $f$ to semi-concavity (which is weaker, as any $L$-smooth function is also $L$-concave).

[3] In the original CCCP [65], $g$ is differentiable, thus the update becomes $-\nabla f(\mathbf{x}^k) = \nabla g(\mathbf{x}^{k+1})$.

[4] Note that $g$ is still assumed to be convex, so that the subproblem (10) can be solved to global optimality.

Third, we also consider much weaker stepsize schemes such as *constant* or *non-summable* ones. We show that for either concave $f$ or strongly-convex $g$, a better rate of convergence $\mathcal{O}(1/k)$ can be achieved, even under the (weak) constant stepsize. It should be noted that our results are new.

All the results in this section are stated under the following assumptions, where $L_f$ and $\sigma_g$ are non-negative constants and $\|\cdot\|$ denotes the $\ell_2$ norm. Their proofs are given in Appendix A.4.

**Assumption 1.** *$f$ is differentiable and $L_f$-semi-concave (i.e., $f(\mathbf{x}) - \frac{L_f}{2}\|\mathbf{x}\|^2$ is concave) on $\operatorname{dom} f$, which is assumed to be open and convex. When $L_f = 0$, $f$ is concave.*

**Assumption 2.** *$g$ is proper, closed, and $\sigma_g$-strongly-convex (i.e., $g(\mathbf{x}) - \frac{\sigma_g}{2}\|\mathbf{x}\|^2$ is convex), and $\operatorname{dom} g \subseteq \operatorname{dom} f$ is compact. When $\sigma_g > 0$, $g$ is strongly convex.*

Denote by $\mathbf{p}_{\mathbf{x}}$ a solution of $\min_{\mathbf{p}} \{\langle \nabla f(\mathbf{x}), \mathbf{p}\rangle + g(\mathbf{p})\}$ and let $\mathbf{p}^k = \mathbf{p}_{\mathbf{x}^k}$. The following quantity, called the *conditional gradient norm*, will serve as an optimality measure:

$$S(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{p}_{\mathbf{x}}\rangle + g(\mathbf{x}) - g(\mathbf{p}_{\mathbf{x}}). \tag{13}$$

**Lemma 1.** *$S(\mathbf{x}) \geq \frac{\sigma_g}{2}\|\mathbf{x} - \mathbf{p}_{\mathbf{x}}\|^2 \ \forall \mathbf{x} \in \operatorname{dom} f$, and $S(\mathbf{x}) = 0$ iff $\mathbf{x}$ is a stationary point of (9).*

The following theorem contains our convergence results for the most common stepsize schemes.

**Theorem 1.** *Let $F^*$ be the minimum value of $F$, $\Omega$ be the diameter of $\operatorname{dom} g$, $\Delta_k = F(\mathbf{x}^k) - F^*$, $\omega = \frac{\sigma_g}{L_f + \sigma_g}$, $\rho(\alpha) = \alpha \min\left\{1, 2 - \frac{\alpha}{\omega}\right\}$, $\eta(\alpha) = \frac{1}{2}\left[(L_f + \sigma_g)\alpha - \sigma_g\right]$, and $\mu = \sqrt{2L_f\Delta_0}$. For any $k \geq 0$, we have $\min_{0\leq i\leq k} S(\mathbf{x}^i) \leq B_k$, where the bound $B_k$ is given as follows:*

| | constant stepsize $\alpha_k = \alpha > 0 \ \forall k$ | constant step length $\alpha_k = \frac{\alpha}{\|\mathbf{p}^k - \mathbf{x}^k\|} \ \forall k$ | non-summable $\sum_{k=0}^{+\infty} \alpha_k = \infty$ | adaptive and line search (14) |
|---|---|---|---|---|
| convex $g$ | $\frac{\Delta_0}{\alpha(k+1)} + \frac{L_f\Omega^2\alpha}{2}$ | $\frac{\Delta_0\Omega}{\alpha(k+1)} + \frac{L_f\Omega\alpha}{2}$ | $\frac{\Delta_0 + \frac{L_f\Omega^2}{2}\sum_{i=0}^k \alpha_i^2}{\sum_{i=0}^k \alpha_i}$ | $\max\left(\frac{2\Delta_0}{k+1}, \frac{\mu\Omega}{\sqrt{k+1}}\right)$ |
| strongly convex $g$ | $\frac{\Delta_0}{\alpha(k+1)} + \eta(\alpha)\Omega^2 \ \forall \alpha \geq 2\omega$<br>$\frac{\Delta_0}{\rho(\alpha)(k+1)} \quad \forall \alpha < 2\omega$ | $\left(\frac{\Delta_0}{\alpha\sqrt{2\sigma_g}(k+1)} + \frac{(L_f+\sigma_g)\alpha}{2\sqrt{2\sigma_g}}\right)^2$ | $\frac{\Delta_{k(\omega)}}{\sum_{i=k(\omega)}^k \alpha_i}$ | $\frac{\Delta_0}{\omega(k+1)}$ |
| concave $f$ | $\frac{\Delta_0}{\alpha(k+1)}$ | $\frac{\Delta_0\Omega}{\alpha(k+1)}$ | $\frac{\Delta_0}{\sum_{i=0}^k \alpha_i}$ | $\frac{2\Delta_0}{k+1}$ |

*In the above, $k(\omega) = \min\{k : \alpha_i < 2\omega \ \forall i \geq k\}$, with further assumption that $\lim_{k\to\infty} \alpha_k = 0$ for (jointly) non-concave $f$ and non-summable $\alpha_k$. The adaptive and line-search stepsizes are given by:*

$$\alpha_k = \min\left\{1, \frac{1}{L_f + \sigma_g}\left(\frac{S(\mathbf{x}^k)}{\|\mathbf{p}^k - \mathbf{x}^k\|^2} + \frac{\sigma_g}{2}\right)\right\}, \qquad \alpha_k = \operatorname*{argmin}_{\alpha\in[0,1]} F(\mathbf{x}^k + \alpha(\mathbf{p}^k - \mathbf{x}^k)). \tag{14}$$

*For the non-highlighted cases, we have $\lim_{k\to\infty} S(\mathbf{x}^k) = 0$ and thus any limit point of $\{\mathbf{x}^k\}$ is a stationary point of (9). The algorithm is guaranteed to reach an approximate one in the other cases.*

Due to space constraints, we give a brief discussion and refer to the appendix (§A.4 and Table 4) for far more detailed results and analyses. Theorem 1 shows that $S(\mathbf{x}^k)$ converges to 0 with rate of convergence $\mathcal{O}(1/\sqrt{k})$ for adaptive or line-search stepsizes Beck [7]. This rate can be improved to $\mathcal{O}(1/k)$ even for weaker stepsize schemes, if $f$ is concave or $g$ is strongly convex. In particular, convergence is guaranteed when $f$ is concave, for which the best bound is obtained when $\alpha_k = 1 \forall k$, which explains the default unit stepsize in the concave-convex procedure [65] (see §3.4). Convergence is also guaranteed for the (diminishing) non-summable scheme (which includes common stepsizes such as $\alpha_k = 2/(k+2)$ or $\alpha_k = 1/\sqrt{k}$), but the rate depends on the rate of divergence of $\sum_{i=0}^k \alpha_i$.

For all the instantiations of regularized Frank-Wolfe presented in §3.3 and §3.4, it is easy to check that Assumptions 1 and 2 are satisfied, as our energy function is quadratic while $\mathcal{X}$ is closed and convex. In addition, the regularizers in most of them ($\ell_2$ or Entropy Frank-Wolfe, including mean field) are strongly convex, thus we would expect a rate of convergence of at least $\mathcal{O}(1/k)$ in practice for these algorithms, under the adaptive, line search, or (suitable) constant stepsizes.

### 4.2 Tightness of the relaxation

We have seen that regularized Frank-Wolfe (Algorithm 1) minimizes a modified continuous energy. It is thus reasonable to ask whether doing so also minimizes the original discrete energy (which is

the main objective). In this section, we partially answer this question by providing some tightness guarantee for this regularized continuous relaxation. Our analysis is quite general and also includes several existing tightness results [8, 39, 55] as special cases. All proofs can be found in Appendix A.5.

The last step in Algorithm 1 consists in converting $\mathbf{x}$ to a discrete solution. We consider two such rounding schemes. The simplest one is perhaps **nearest rounding**, which assigns each node $i$ with the label $s_i \in \operatorname{argmax}_{t \in \mathcal{S}} x_{it}$. Intuitively, this sets $\mathbf{x}_i$ to the nearest vertex of the simplex $\left\{\mathbf{x}_i \in \mathbb{R}_+^d : \mathbf{1}^\top \mathbf{x}_i = 1\right\}$. The second scheme, called **BCD rounding** [39, 55], consists in minimizing $E(\mathbf{x})$ over $\mathbf{x}_i$ while keeping all $\mathbf{x}_j$ ($j \neq i$) fixed (i.e., block coordinate descent), which amounts to iteratively assigning each node $i$ with label $s_i \in \operatorname{argmin}_{s \in \mathcal{S}} \left\{\theta_i(s) + \sum_{j \in \mathcal{N}_i} \sum_{t \in \mathcal{S}} \theta_{ij}(s,t) x_{jt}\right\}$. In practice, we only use nearest rounding because BCD rounding is too expensive for dense graphs. However, an important property of the latter is that it does not increase the energy, which is useful for our theoretical analysis. The following theorem provides an additive bound on the energy.

**Theorem 2.** *Let $\mathbf{x}_r^*$ be a global minimum of $E_r(\mathbf{x}) = E(\mathbf{x}) + r(\mathbf{x})$ over $\mathcal{X}$, $\bar{\mathbf{x}}_r^*$ be the discrete solution rounded from $\mathbf{x}_r^*$, and $E^*$ be the minimum discrete energy. Assume that $r(\mathbf{x})$ is bounded:*[5] *$m \leq r(\mathbf{x}) \leq M \ \forall \mathbf{x} \in \mathcal{X}$. We have $E^* \leq E(\bar{\mathbf{x}}_r^*) \leq E^* + M - m + C$, where $C = \sqrt{n\left(1 - \frac{1}{d}\right)}\left(\|\mathbf{u}\|_2 + \sqrt{n}\,\|\mathbf{P}\|_2\right)$ for nearest rounding and $C = 0$ for BCD rounding.*

Let us derive the energy BCD bound for some particular cases (see §A.5 for details). Obviously with no regularization ($r = 0$), we have $M = m = 0$ and thus $E(\bar{\mathbf{x}}_r^*) \leq E^* \leq E(\bar{\mathbf{x}}_r^*)$, yielding $E(\bar{\mathbf{x}}_r^*) = E^*$, i.e., the relaxation is tight. We have thus recovered a previously known result [8, 39, 55]. For $r(\mathbf{x}) = -\mathbf{c}^\top \mathbf{x} + \mathbf{x}^\top \operatorname{diag}(\mathbf{c}) \mathbf{x}$ with $\mathbf{c} \geq \mathbf{0}$, we have $M = 0$ and $m = -\frac{1}{4}\mathbf{1}^\top \mathbf{c}$, which recovers exactly the additive bound given by Ravikumar and Lafferty [55] for the convex QP relaxation. For the $\ell_2$ regularizer $r(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}\|_2^2$, we have $M = \frac{\lambda n}{2}$ and $m = \frac{\lambda n}{2d}$, thus we obtain a bound of $\frac{\lambda n}{2}\left(1 - \frac{1}{d}\right)$. For the entropy regularizer $r(\mathbf{x}) = -\lambda H(\mathbf{x})$, we have $M = 0$ and $m = -\lambda n \log d$, thus the bound is $\lambda n \log d$, which is worse than the $\ell_2$ bound for any $d \geq 5$.

Note that the bound provided by Theorem 2 is achieved from a *global* minimum of the regularized relaxation. This can be attained in some cases, e.g., when the (convex) regularizer is large enough so that the objective becomes convex. In the general case, however, the algorithm is only guaranteed to reach a stationary point, and the (theoretical) quality of such point remains an open question.

# 5 Experiments

We compare regularized Frank-Wolfe with existing methods on the semantic segmentation task, in terms of both *inference* and *learning* performance. Two variants, namely $\ell_2$ Frank-Wolfe ($\ell_2$**FW**) and Entropy Frank-Wolfe ($e$**FW**) (§3.3), will be compared to the following methods: Mean field (**MF**) [33, 34] (which is considered as our baseline), nonconvex vanilla Frank-Wolfe (**FW**) [39] (§2.3), projected gradient descent (**PGD**) [38, 39], proximal gradient methods (**PGM**) [46], and alternating direction method of multipliers (**ADMM**) [39]. Performance of *convex* vanilla Frank-Wolfe [20] was found to be far inferior to the nonconvex counterpart, and thus excluded from the presentation. Other methods based on CCCP [20] or LP relaxation [1, 20] are also excluded due to their sophisticated implementations. For all methods, we set the initial solution to $\mathbf{x}^0 = \operatorname{softmax}(-\mathbf{u})$ as previously done [33]. Details on implementation, running time, and memory footprints are given in Appendix B.

## 5.1 Experimental setup

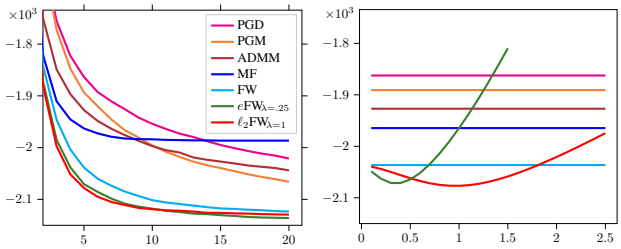
Our segmentation model is a standard combination of a CNN and a CRF [34, 66] (Appendix B.2). For the CNN part, we consider two strong architectures: DeepLabv3 with ResNet101 backbone [15], and DeepLabv3+ with Xception65 backbone [16]. The CRF part is a fully-connected one [33] in which any pair of pixels $(i, j)$ is an edge with potential $\theta_{ij}(s,t) = \mu(s,t) k(\mathbf{f}_i, \mathbf{f}_j)\ \forall s, t \in \mathcal{S}$, where $\mu : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is called label compatibility function, and $k$ is a Gaussian kernel over image features based on pixel coordinates and colors. The setup of our models are similar to Zheng et al. [66]. We use the Potts compatibility function: $\mu(s,t) = w \mathbb{1}_{[s \neq t]}$ with $w = 1$ for the inference experiments in §5.2, and also for CRF initialization in the learning experiments in §5.3. For all experiments, a

[5] A sufficient condition is $r$ being continuous, as $\mathcal{X}$ is compact.

fully-trained CNN is needed. We follow closely the published recipes [15, 16] for this task. We first pretrain DeepLabv3 and DeepLabv3+ on the COCO dataset [44] and then finetune them on PASCAL VOC (*trainaug*) and Cityscapes (*train*) to obtain similar results to previous work [15, 16] (Table 1, CNN column). Finally, we perform experiments on two popular datasets: (augmented) PASCAL VOC [21] and Cityscapes [18]. Further details are given in Appendix B.

### 5.2 Inference performance

In this section, we compare the performance of regularized Frank-Wolfe against the competing methods in terms of inference. We consider a Potts CRF on top of a CNN, which is the typical setup for using dense CRF in post-processing. Figure 1 shows the *discrete* energy per inference iteration for each method, averaged over the 1449 *val* images of PASCAL VOC, using DeepLabv3+. One can observe that Frank-Wolfe variants completely outperform the other methods. In addition, regularized Frank-Wolfe outperforms all the other methods for a large range of $\lambda$ (Figure 1b).




**(a)** Energy per CRF iteration. **(b)** Per $\lambda$, at the 5$^{\text{th}}$ iteration.

**Figure 1:** CRF energy averaged over 1449 *val* images of PASCAL VOC using DeepLabv3+ and Potts dense CRF. **(a)** Comparison with other methods. **(b)** Energy per $\lambda$ after 5 iterations.

| | | CNN | PGD | PGM | ADMM | MF | FW | $e$FW$_{.7}$ | $e$FW$_{.3}$ | $\ell_2$FW |
|---|---|---|---|---|---|---|---|---|---|---|
| VOC | DL3 | 81.83 | 82.23 | 82.23 | 82.22 | 82.21 | 82.27 | 82.26 | **82.29** | **82.29** |
| | DL3+ | 82.89 | 83.36 | 83.37 | 83.38 | 83.45 | 83.43 | 83.45 | 83.48 | **83.50** |
| CITY | DL3 | 76.73 | 76.88 | 76.86 | 76.95 | 76.97 | 76.86 | 76.99 | 76.99 | **77.03** |
| | DL3+ | 79.55 | 79.64 | 79.63 | **79.66** | 79.63 | 79.64 | 79.65 | **79.66** | **79.66** |

**Table 1:** Performance (mIoU) on the *val* sets, using a Potts dense CRF on top of the pre-trained CNN models. DL means DeepLab.

Table 1 shows the performance on the validation sets of PASCAL VOC and Cityscapes, for a Potts CRF with both DeepLabv3 and DeepLabv3+ as backbone. In this experiment, we run all the methods for 10 iterations. One can observe that $\ell_2$FW achieved the best performance, followed by $e$FW. We should note some inconsistency with respect to the energy results. For example, $e$FW achieved much lower energy than MF, yet the mIoU gap is marginal; also, FW accuracy is slightly worse than MF while the energy is much lower. This can be explained by the fact that the Potts model is not a perfect representation (i.e., lower energy in this model does not necessarily translate to higher accuracy). In the next section, we will see how the methods perform when the CRF parameters are learned from data.

### 5.3 Learning performance

In this section, we evaluate the performance of the methods for joint CNN-CRF end-to-end training with SGD. The CNN is initialized with its fully-trained weights on the corresponding dataset, and the CRF is initialized with the Potts model with random noise added. We train the model for 20 epochs with 5 CRF iterations, using the same poly learning schedule as before. As the CNN has been already fully-trained, we set its learning rate to a small value of 0.0001. For the CRF, we tried 4 different values of initial learning rates $\eta_0 \in \{1.0, 0.1, 0.01, 0.001\}$ and found that 1.0 is too high (training diverges quickly) while 0.001 is too low (slow progress) for all methods. For the remaining learning rates $\{0.1, 0.01\}$, we perform 4 additional trainings for each method (i.e., a total of 5 runs for each configuration).

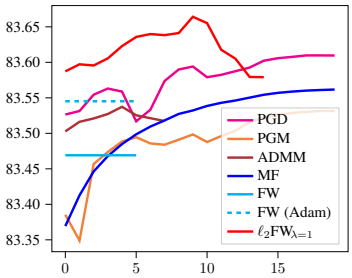



**Figure 2:** *val* mIoU per epoch on PASCAL VOC, showing that vanilla FW fails to learn (zero-gradient issue).

Let us summarize our findings. First, we observe that (vanilla) FW fails to learn. This is illustrated in Figure 2, where we show the validation accuracy per epoch on PASCAL VOC for each method: FW did not make any progress. We tried a different optimizer (Adam [32]) and obtained similar results. This is expected as the gradient in vanilla FW is zero almost everywhere, as previously discussed in §3.1 (see also §A.1). Our second observation is that training is quite unstable for PGD, PGM, ADMM, $e$FW$_{.3}$, and $\ell_2$FW. In particular, $\eta_0 = 0.1$ is still too high for these methods, and even with $\eta_0 = 0.01$, some of the runs produced bad results. By contrast, MF and $e$FW$_{.7}$ are stable for both learning rates, with 0.1 being slightly better. A possible explanation is that PGD, PGM, ADMM, and $\ell_2$FW all

employ a simplex projection step that is not fully differentiable (but only so almost everywhere). For $e$FW$_{.3}$ (which is fully differentiable), we hypothesize that the low regularization makes the problem less "smooth", which may also harm gradient-based training. Finally, with the above training scheme, we observe that none of the CRF methods could improve over the CNN (but rather the opposite) on Cityscapes. We have seen that the Potts CRF was able to achieve some marginal improvements (Table 1), thus it is reasonable to expect even better performance with end-to-end training.

In view of the above observations, we present a simple trick to make CRF training more stable. The idea is to replace the CRF output $\mathbf{x}^*$ with $\frac{1}{2}(\mathbf{x}^* + \mathbf{x}^0)$, where we recall that the initialization $\mathbf{x}^0$ is the $\mathrm{softmax}$ of the CNN logits. Intuitively, this adds a skip connection from the CNN to the CRF output in the computation graph, which makes the gradient of the loss propagate directly to the CNN. We found that this trick also slightly improves $e$FW$_{.7}$, but has a negative effect on MF. Therefore, it is applied to all methods except MF. Finally, as Cityscapes requires a very high number of epochs, we set this value to $100$. Also because training on Cityscapes requires a lot more computing resources, we only perform a single run on DeepLabv3+. The results are presented in Table 2. Again, $\ell_2$FW consistently achieves the best results. Interestingly, while $e$FW$_{.3}$ achieved similar performance to $\ell_2$FW in terms of energy minimization (Figure 1 and Table 1), it achieved the worse performance in joint training. Compared to Table 1, one can observe that joint training produced much larger improvements over the CNNs, up to $2.25$ points on PASCAL VOC and $0.4$ points on Cityscapes.

| | | CNN | PGD | PGM | ADMM | MF | $e$FW$_{.7}$ | $e$FW$_{.3}$ | $\ell_2$FW |
|---|---|---|---|---|---|---|---|---|---|
| VOC | DL3 | 81.83 | 83.69 $\pm 0.20$ | **83.75** $\pm 0.23$ | 83.68 $\pm 0.06$ | 83.69 $\pm 0.10$ | 83.50 $\pm 0.10$ | 83.25 $\pm 0.20$ | **83.75** $\pm 0.13$ |
| | DL3+ | 82.89 | 84.82 $\pm 0.23$ | 84.79 $\pm 0.20$ | 84.83 $\pm 0.06$ | 84.87 $\pm 0.17$ | 84.64 $\pm 0.23$ | 84.50 $\pm 0.16$ | **85.14** $\pm 0.09$ |
| CITY | DL3+ | 79.55 | 79.80 | 79.62 | 79.62 | 79.74 | 79.70 | 79.58 | **79.95** |

**Table 2:** Performance on the *val* sets for joint training. For PASCAL VOC, we report the mean and standard deviation from $5$ runs.

**Performance on the test sets** We select the best performing method (DeepLabv3+ with $\ell_2$FW CRF) for evaluation on the test sets. For PASCAL VOC, we further train our model on the union of the *train* and *val* subsets for $50$ epochs. For Cityscapes, we further train $200$ epochs on *train* and *train_extra*, using the high-quality annotations provided by Tao et al. [62] (for *train_extra*). At the $150^{\text{th}}$ epoch, we replace *train_extra* with *val*. For this fine-tuning step, learning rates were set to $0.001$ for CNN and $0.1$ for CRF. For prediction, we apply test time augmentation including left-right flipping and multi-scales. For reference, we train DeepLabv3+ alone using the same recipes. Table 3 shows that we were able to closely match the performance reported by Chen et al. [16]. Adding the $\ell_2$FW CRF yields an improvement of $0.4$ points on PASCAL VOC. Unfortunately we only observe a marginal improvement on Cityscapes.

| Model | VOC | CITY |
|---|---|---|
| DeepLabv3+ [16] | 87.8 | 82.1 |
| DeepLabv3+ (this work) | 87.6 | 83.5 |
| DeepLabv3+ with $\ell_2$FW CRF | **88.0** | **83.6** |

**Table 3:** Performance on the *test* sets. Submission URLs are given in Appendix C.2.

## 6 Discussion & conclusion

**Why does it work?** Theoretically, all the methods in §5 should reach a stationary point, so why is one better than the other? In fact, Figure 1a only shows that Frank-Wolfe based methods work better than the other *in the first few iterations*, but not necessarily in a later stage. Indeed, the same conclusion no longer holds after $100$ iterations (see §C.1), but this long regime is not interesting to us as only a few (typically $5$) iterations should be used to avoid the vanishing/exploding gradient problems [66]. As to why Frank-Wolfe achieves lower energy in the early stage, we hypothesize that this could be due to the discreteness of its iterates (7). With small $\lambda$, the solution by regularized Frank-Wolfe should be close to the vanilla one, and thus also benefits from this property. It is important to note that the benefit of regularized Frank-Wolfe does not lie in the extra (sometimes small) energy improvement over vanilla Frank-Wolfe, but in its ability to seamlessly solve the zero-gradient issue.

**How to tune $\lambda$?** We found that similar curves to Figure 3b can be obtained using a small random subset (e.g., $10$ samples) of the data, which suggests a quick way of tuning $\lambda$ by random subsampling. In practice, this step takes only a few seconds, which is negligible in most training scenarios.

**Limitations** While one variant of regularized Frank-Wolfe ($\ell_2$FW) consistently achieves the best results, the difference compared to the other methods is sometimes small. In addition, the improvement of dense CRFs over CNNs is marginal on the Cityscapes test set. Nevertheless, we hope the encouraging results on PASCAL VOC could attract interest from the community in CRF research, potentially leading to creative ways of overcoming these limitations.

# References


[1] Thalaiyasingam Ajanthan, Alban Desmaison, Rudy Bunel, Mathieu Salzmann, Philip HS Torr, and M Pawan Kumar. Efficient linear programming for dense crfs. In *Computer Vision and Pattern Recognition*, 2017.

[2] Thalaiyasingam Ajanthan, Puneet K Dokania, Richard Hartley, and Philip HS Torr. Proximal mean-field for neural network quantization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4871–4880, 2019.

[3] Brandon Amos, Vladlen Koltun, and J Zico Kolter. The limited multi-label projection layer. *arXiv preprint arXiv:1906.08707*, 2019.

[4] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision*, pages 524–540. Springer, 2016.

[5] Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

[6] Pierre Baqué, Timur Bagautdinov, François Fleuret, and Pascal Fua. Principled parallel mean-field inference for discrete random fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5848–5857, 2016.

[7] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[8] Marc Berthod. Definition of a consistent labeling as a global extremum. In *International Conference on Pattern Recognition*, pages 339–341, 1982.

[9] Marc Berthod, Gérard Giraudon, and Jean Paul Stromboni. Deterministic pseudo-annealing: Optimization in markov-random-fields an application to pixel classification. In *European Conference on Computer Vision*, pages 67–71. Springer, 1992.

[10] Kristian Bredies, Dirk A Lorenz, and Peter Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, Advanced online publication, 2007. doi: 10.1007/s10589-007-9083-3.

[11] Kristian Bredies, Dirk A Lorenz, and Peter Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*, 42(2):173–193, 2009.

[12] Siddhartha Chandra and Iasonas Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *European conference on computer vision*, pages 402–418. Springer, 2016.

[13] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.

[14] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[15] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[16] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[17] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[19] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.

[20] Alban Desmaison, Rudy Bunel, Pushmeet Kohli, Philip HS Torr, and M Pawan Kumar. Efficient continuous relaxations for dense crf. In *European Conference on Computer Vision*, pages 818–833. Springer, 2016.

[21] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[22] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[23] Robert M Freund and Paul Grigas. New analysis and results for the frank–wolfe method. *Mathematical Programming*, 155(1-2):199–230, 2016.

[24] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

[25] Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Mathematical Programming*, 152(1-2):75–112, 2015.

[26] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[29] Xu Hu and Guillaume Obozinski. Sdca-powered inexact dual augmented lagrangian method for fast CRF learning. In *AISTATS*, volume 84 of *Proceedings of Machine Learning Research*, pages 988–997. PMLR, 2018.

[30] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th international conference on machine learning*, pages 427–435, 2013.

[31] Vladimir Jojic, Stephen Gould, and Daphne Koller. Accelerated dual decomposition for map inference. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 503–510, 2010.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[34] Philipp Krähenbühl and Vladlen Koltun. Parameter learning and convergent inference for dense random fields. In *International Conference on Machine Learning*, pages 513–521, 2013.

[35] Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants. In *Advances in neural information processing systems*, pages 496–504, 2015.

[36] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, 2001. ISBN 1-55860-778-1.

[37] Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

[38] Måns Larsson, Anurag Arnab, Fredrik Kahl, Shuai Zheng, and Philip Torr. A projected gradient descent method for crf inference allowing end-to-end training of arbitrary pairwise potentials. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 564–579. Springer, 2017.

[39] D Khuê Lê-Huu and Nikos Paragios. Continuous relaxation of map inference: A nonconvex perspective. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[40] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1 (4):541–551, 1989.

[41] Xiaoxiao Li, Ziwei Liu, Ping Luo, Chen Change Loy, and Xiaoou Tang. Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3193–3202, 2017.

[42] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2016.

[43] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[45] Seppo Linnainmaa. The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), University of Helsinki*, pages 6–7, 1970.

[46] Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[47] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE international conference on computer vision*, pages 1377–1385, 2015.

[48] Julien Mairal. Optimization with first-order surrogate functions. In *International Conference on Machine Learning*, pages 783–791, 2013.

[49] Ofer Meshi, Mehrdad Mahdavi, and Alex Schwing. Smooth and strong: Map inference with linear convergence. In *Advances in Neural Information Processing Systems*, pages 298–306, 2015.

[50] Hisashi Mine and Masao Fukushima. A minimization method for the sum of a convex function and a continuously differentiable function. *Journal of Optimization Theory and Applications*, 33(1):9–23, 1981.

[51] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1): 127–152, 2005. ISSN 0025-5610.

[52] Vlad Niculae and Mathieu Blondel. A regularized framework for sparse and structured neural attention. In *Advances in neural information processing systems*, pages 3338–3348, 2017.

[53] Giorgio Parisi. *Statistical field theory*. Addison-Wesley, 1988.

[54] Marin Vlastelica Pogancic, Anselm Paulus, Vít Musil, Georg Martius, and Michal Rolinek. Differentiation of blackbox combinatorial solvers. In *ICLR*. OpenReview.net, 2020.

[55] Pradeep Ravikumar and John Lafferty. Quadratic programming relaxations for metric labeling and markov random field map estimation. In *Proceedings of the 23rd international conference on Machine learning*, pages 737–744. ACM, 2006.

[56] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[57] Alexander G Schwing and Raquel Urtasun. Fully connected deep structured networks. *arXiv preprint arXiv:1503.02351*, 2015.

[58] Falong Shen, Rui Gan, Shuicheng Yan, and Gang Zeng. Semantic segmentation via structured patch prediction, context crf and guidance crf. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1953–1961, 2017.

[59] Solomon Eyal Shimony. Finding maps for belief networks is np-hard. *Artificial Intelligence*, 68(2): 399–410, 1994.

[60] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *ICML-31st International Conference on Machine Learning*, volume 32, pages 1611–1619. JMLR, 2014.

[61] Haiming Sun, Di Xie, and Shiliang Pu. Mixed context networks for semantic segmentation. *arXiv preprint arXiv:1610.05854*, 2016.

[62] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020.

[63] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.07122.

[64] Yaoliang Yu, Xinhua Zhang, and Dale Schuurmans. Generalized conditional gradient for sparse estimation. *The Journal of Machine Learning Research*, 18(1):5279–5324, 2017.

[65] Alan L Yuille and Anand Rangarajan. The concave-convex procedure (cccp). In *Advances in neural information processing systems*, pages 1033–1040, 2002.

[66] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code will be available at publication time.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Some details in the main content and full details in the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In the appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] We use standard datasets freely available for academic research. They are well-known so we did not mention their licenses.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] We will release our pre-trained models at publication time.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]