Supplementary Material for
SDCA-Powered Inexact Dual Augmented Lagrangian Method for Fast CRF Learning

A Loss-Augmented CRF

In order to extend our learning formulation so as to encompass as well max-margin structured learning (i.e., structured SVM) in additional to maximum likelihood learning, we show in this section that our formulation can be generalized to cover the loss-augmented CRF learning introduced by Pletscher et al. (2010) and Hazan and Urtasun (2010).

The loss-augmented CRF $p_\gamma(y \mid y^\star, x)$ is an extension of the standard CRF with additional user-defined loss functions $\ell_c(y_c^\star, y_c)$ for all cliques and an extra temperature hyperparameter $\gamma \in (0, +\infty)$. We introduce a modified natural parameter $\eta_\ell(w) := \eta(w) + \ell$ (similarly we have θ_ℓ) that includes the loss term $\ell = \big[[\ell_c(y_c^\star, y_c) \colon y_c \in \mathcal{Y}_c] \colon c \in \mathcal{C}\big]$. The density function of the loss-augmented CRF then takes the form

$$p_\gamma(y \mid y^\star, x; w) = \exp\Big(\langle \eta_\ell(w)/\gamma, T(y)\rangle - F(\eta_\ell(w)/\gamma)\Big). \tag{1}$$

A justification for the form of the loss-augmented CRF is based on a rationale that distinguishes the label to predict y (which is essentially true unknown label) from the label provided by the annotation $y^\star$. The assumption made is then that, given y_c, the annotation $y_c^\star$ is independent of x and $y_{c'}$ for $c' \neq c$. This entails that $p(y, y^\star|x) = p_\gamma(y \mid y^\star, x) \propto p(y \mid y^\star) p_\gamma(y \mid x)$, which yields the above form for $p_\gamma(y \mid y^\star, x; w)$ by Bayes' rule for $p(y \mid y^\star) \propto \exp(\sum_{c\in\mathcal{C}} \ell_c(y_c^\star, y_c))$.

For learning, we use a rescaled maximum likelihood objective (i.e., multiplied by γ) of the form

$$\min_w \gamma F\Big(\tfrac{1}{\gamma}\theta_\ell(w)\Big) + \frac{\lambda}{2}\|w\|_2^2, \tag{2}$$

with which we can see γ only affects the entropy term in the variational representation of F, thus it plays a role to determine the learning regime. When $\gamma \to 0$, we retrieve a max-margin formulation for structured output learning, since the corresponding variational problem based on Fenchel duality is

$$\min_w \max_{\mu\in\mathcal{M}} \langle \mu, \theta_\ell\rangle + \frac{\lambda}{2}\|w\|_2^2. \tag{3}$$

Note that this is identical to the linear programming relaxation of the structured SVM formulation studied by Meshi et al. (2010).

It is also possible to retrieve the maximum likelihood regime by making a change of variable: $w' = w/\gamma$. Then, (2) becomes

$$\min_{w'} F\Big(\theta(w') + \frac{1}{\gamma}\ell\Big) + \frac{\lambda\gamma}{2}\|w'\|_2^2. \tag{4}$$

Increasing γ decreases the effect of the loss term and simultaneously increases the effect of the regularization. The maximum likelihood regime is thus retrieved by letting $\gamma \to +\infty$ and $\lambda \to 0$.

B Derivations of dual, and relaxed primal and dual objectives

In this section, we derive the dual objective $D(\mu)$ of $P(w)$. Given the augmented Lagrangian $D_\rho(\mu, \xi)$, we first introduce a relaxed primal $\tilde{P}_\rho(w, \delta, \xi)$ involving a new primal variable δ whose components can be interpreted as messages exchanged between cliques in the context of marginal inference via message-passing algorithms. The partial minimization with respect to δ then yields the corresponding primal of $D_\rho(\mu, \xi)$ with respect to μ for a fixed ξ: $P_\rho(w, \xi) := \min_\delta \tilde{P}_\rho(w, \delta, \xi)$, which can be interpreted as a Moreau-Yoshida smoothing of the original objective $P_\rho(w)$.

B.1 Derivation of the dual objective $D(\mu)$

Given that $\theta_\ell(w) = \Psi^\intercal w + \ell$ and introducing the Fenchel conjugate of $F_\mathcal{L}$, we have

$$\begin{aligned} P(w) &= \gamma F_\mathcal{L}\Big(\frac{1}{\gamma}\theta_\ell(w)\Big) + \frac{\lambda}{2}\|w\|_2^2 \\ &= \max_{\mu \in \mathcal{L}} \Big[\langle \Psi^\intercal w + \ell, \mu\rangle - \gamma F_\mathcal{L}^*(\mu)\Big] + \frac{\lambda}{2}\|w\|_2^2. \end{aligned}$$

Given that the local polytope constraints are defined by linear inequalities, weak Slater constraint qualification are satisfied, so that strong duality holds and an equivalent dual problem in μ is obtained by switching the order of $\min_w$ and $\max_\mu$:

$$\begin{aligned} D(\mu) &= \langle \ell, \mu\rangle - \gamma F_\mathcal{L}^*(\mu) + \min_w \Big[\langle \Psi^\intercal w, \mu\rangle + \frac{\lambda}{2}\|w\|_2^2\Big] \\ &= \langle \ell, \mu\rangle - \gamma F_\mathcal{L}^*(\mu) - \lambda \max_w \Big[-\frac{1}{\lambda}\langle \Psi\mu, w\rangle - \frac{1}{2}\|w\|_2^2\Big] \\ &= \langle \ell, \mu\rangle - \gamma F_\mathcal{L}^*(\mu) - \frac{1}{2\lambda}\|\Psi\mu\|_2^2. \end{aligned}$$

B.2 Derivation of an extended primal $\tilde{P}_\rho(w, \delta, \xi)$

Proposition 4. *For a fixed ξ, the primal objective function of $D_\rho(\mu, \xi)$ takes the form*

$$P_\rho(w, \xi) := \min_\delta \Big[\tilde{P}_\rho(w, \delta, \xi) := \gamma F_\mathcal{I}\Big(\frac{\theta(w) + A^\intercal\delta}{\gamma}\Big) + \frac{\lambda}{2}\|w\|_2^2 + \frac{\rho}{2}\|\delta - \xi\|^2\Big].$$

Proof. Clearly, we have $D(\mu) = \min_\xi D_\rho(\mu, \xi)$. For a fixed value of ξ, consider the Lagrangian

$$L_{\rho,\xi}(\mu, \nu, \nu', w, \delta) = \langle \ell, \mu\rangle - \gamma F_\mathcal{I}^*(\mu) - \frac{1}{2\lambda}\|\nu\|^2 - \frac{1}{2\rho}\|\nu'\|^2 + \langle \xi, \nu'\rangle + \langle w, \Psi\mu - \nu\rangle + \langle \delta, A\mu - \nu'\rangle;$$

Then clearly $\min_{w,\delta} L_{\rho,\xi}(\mu, \nu, \nu'; w, \delta) = D_\rho(\mu, \xi)$. We compute the associated primal as

$$\begin{aligned} &\tilde{P}_\rho(w, \delta, \xi) = \max_{\mu,\nu,\nu'} L_{\rho,\xi}(\mu, \nu, \nu', w, \delta) \\ &= \max_u \Big[\langle \mu, \ell + \Psi^\intercal w + A^\intercal\delta\rangle - \gamma F_\mathcal{I}^*(\mu)\Big] + \max_\nu \Big[\langle \nu, -w\rangle - \frac{1}{2\lambda}\|\nu\|^2\Big] + \max_{\nu'} \Big[\langle \nu', \xi - \delta\rangle - \frac{1}{2\rho}\|\nu'\|^2\Big], \end{aligned}$$

which yields the desired form of $P_\rho(w, \xi) = \min_\delta \tilde{P}_\rho(w, \delta, \xi)$ upon expliciting Fenchel conjugates. □

Proposition 5.

$$\min_\delta F_\mathcal{I}\big(\tfrac{1}{\gamma}(\theta(w) + A^\intercal\delta)\big) = F_\mathcal{L}\big(\tfrac{1}{\gamma}\theta(w)\big) \qquad \textit{and} \qquad \min_{\xi,\delta} \tilde{P}_\rho(w, \delta, \xi) = P(w).$$

Proof. We have

$$\begin{aligned}\min_{\delta} F_{\mathcal{I}}\big(\tfrac{1}{\gamma}(\theta(w)+A^{\intercal}\delta)\big) &= \min_{\delta}\max_{\mu}\Big(\frac{1}{\gamma}\langle\theta(w)+A^{\intercal}\delta,\mu\rangle + H_{\text{Approx}}(\mu) - \iota_{\mathcal{I}}(\mu)\Big)\\ &= \max_{\mu}\Big(\frac{1}{\gamma}\langle\theta(w),\mu\rangle + H_{\text{Approx}}(\mu) - \iota_{\mathcal{I}}(\mu) - \iota_{\{A\mu=0\}}\Big)\\ &= F_{\mathcal{L}}\big(\tfrac{1}{\gamma}\theta(w)\big),\end{aligned}$$

where the second equality follows by exchanging minimization and maximization (strong duality holds by Slater's conditions) and minimizing with respect to δ.

To show that $\min_{\xi,\delta}\tilde{P}_{\rho}(w,\delta,\xi) = P(w)$, it is easy to minimize over ξ first, which cancels out the term $\frac{\rho}{2}\|\delta-\xi\|^2$ by setting $\xi=\delta$. Then, δ only appears in $F_{\mathcal{I}}$ and the result follows from the first result. □

B.3 Interpretation as Moreau-Yosida smoothing

To understand the structure of $P_{\rho}(w,\xi)$, we shall look at $\tilde{P}_{\rho}(w,\delta,\xi)$. One may be interested in where does δ comes from? In fact, forming the Lagrangian of $\min_w P(w)$ with Lagrangian multiplier δ corresponding to the marginalization constraint $A\mu=0$, we see that

$$L(w,\delta,\mu) := \langle\theta_{\ell}(w),\mu\rangle - \gamma F^{*}_{\mathcal{I}}(\mu) + \frac{\lambda}{2}\|w\|_2^2 + \langle\delta, A\mu\rangle.$$

Recall that the Moreau-Yosida regularization of a function f is defined as the infimal convolution

$$M_{\rho f}(x) = \min_{z}\Big[f(z) + \frac{\rho}{2}\|z-x\|^2\Big].$$

Both $P_{\rho}(w,\xi)$ and $D_{\rho}(\mu,\xi)$ have a nice interpretation in terms of the Lagrangian L and Moreau-Yosida regularization. Note that the Moreau-Yosida regularization admits the same optimum as the original function, and that it is smooth even when the original function is not. It is furthermore $\frac{\gamma\rho}{\gamma+\rho}$-strongly convex if the original function is γ-strongly convex.

Proposition 6. *$P_{\rho}(w,\xi)$ and $D_{\rho}(\mu,\xi)$ are respectively the Moreau-Yosida regularizations of $L_{\mu^{\star}}: w,\delta\mapsto\max_{\mu}L(w,\delta,\mu)$ and $L_{w^{\star}}:\mu,\delta\mapsto\min_{w}L(w,\delta,\mu)$ about δ. that is*

$$\begin{aligned}P_{\rho}(w,\xi) &= M_{\rho L_{\mu^{\star}}}(w,\xi) = \min_{\delta}\Big[\max_{\mu}L(w,\delta,\mu) + \frac{\rho}{2}\|\delta-\xi\|_2^2\Big]\\ D_{\rho}(\mu,\xi) &= M_{\rho L_{w^{\star}}}(\mu,\xi) = \min_{\delta}\Big[\min_{w}L(w,\delta,\mu) + \frac{\rho}{2}\|\delta-\xi\|_2^2\Big].\end{aligned}$$

Proof. For $P_{\rho}(w,\xi)$, note that $\max_{\mu}L(w,\delta,\mu)\equiv\gamma F_{\mathcal{I}}\Big(\frac{\theta(w)+A^{\intercal}\delta}{\gamma}\Big)+\frac{\lambda}{2}\|w\|_2^2$. The equivalent form is immediately derived from Proposition 4.

For $D_{\rho}(\mu,\xi)$, note that $\min_w L(w,\delta,\mu)\equiv\langle\theta,\mu\rangle-\gamma F^{*}_{\mathcal{I}}(\mu)-\frac{1}{2\lambda}\|\Psi\mu\|^2+\langle\delta,A\mu\rangle$, and $\min_{\delta}\langle\delta,A\mu\rangle+\frac{\rho}{2}\|\delta-\xi\|^2\equiv\langle\xi,A\mu\rangle-\frac{1}{2\rho}\|A\mu\|_2^2$. Thus, the equivalence holds.

□

Note that the penalty formulation corresponds to a special case of $\tilde{P}_{\rho}(w,\delta,\xi)$ and $D_{\rho}(\mu,\xi)$ with $\xi=0$. It introduces an additional term $\frac{\rho}{2}\|\delta\|^2$, thus making the primal strongly convex with respect to δ and the dual smoother in μ. This effect is similar to that of using Moreau-Yosida smoothing. However, the additional term $\frac{\rho}{2}\|\delta\|^2$ will never vanish, so $A\mu=0$ will never be satisfied. The more $A\mu=0$ is violated, the less the structure of CRF will be perserved.

B.4 Duality gaps and representer theorem

Besides, if we define $\text{gap}(w,\delta,\mu,\xi) := \tilde{P}_\rho(w,\delta,\xi) - D_\rho(\mu,\xi)$ as an upper-bound estimate of the duality gap $P_\rho(w,\xi) - D_\rho(\mu,\xi)$, specifically

$$\begin{aligned}\text{gap}(w,\delta,\mu,\xi) = &\Big[\gamma F_{\mathcal{I}}\big(\tfrac{1}{\gamma}\theta_\ell(w) + A^\intercal\delta\big) + \gamma F^*_{\mathcal{I}}(\mu) - \langle\theta_\ell(w) + A^\intercal\delta, \mu\rangle\Big] \\ &+ \Big[\frac{\lambda}{2}\|w\|^2 + \frac{1}{2\lambda}\|\Psi\mu\|^2 - \langle -w, \Psi\mu\rangle\Big] + \Big[\frac{\rho}{2}\|\xi-\delta\|^2 + \frac{1}{2\rho}\|A\mu\|^2 - \langle\xi-\delta, A\mu\rangle\Big],\end{aligned}$$

we can see that the recovered w and δ by the optimality condition make the 2nd and 3rd term of $\text{gap}(w,\delta,\mu,\xi)$ disappear. We will see later this is important in designing the algorithm to solve $\max_\mu D_\rho(\mu,\xi)$.

Finally, we give a rough picture of all the quantities that we introduced in this section, which can be easily derived from Proposition 6.

Corollary 5. *The relations between D, D_ρ, P and P_ρ could be summarized as*

$$\begin{aligned}&D(\mu) \le D_\rho(\mu,\xi) \le P_\rho(w,\xi);\\ &D_\rho(\mu) \le P(w) \le P_\rho(w,\xi) \le \tilde{P}_\rho(w,\delta,\xi);\\ &\max_\mu \min_\xi D_\rho(\mu,\xi) \le \min_w P(w),\end{aligned}$$

with equalities hold for the saddle point $(\mu^\star, w^\star, \xi^\star)$. Moreover, the first-order optimality conditions are given as

$$w^\star = -\frac{1}{\lambda}\Psi\mu^\star, \quad \delta^\star = \xi^\star - \frac{1}{\rho}A\mu^\star \tag{5}$$

Proof. By constructions, $D(\mu) = \min_\xi D_\rho(\mu,\xi) \le D_\rho(\mu,\xi)$ and $P_\rho(w,\xi) = \min_\delta \tilde{P}_\rho(w,\delta,\xi) \le \tilde{P}_\rho(w,\delta,\xi)$. Other inequalities are the consequences of Proposition 6 and the min-max inequality. Since the strong duality holds (Slater conditions satisfied and the problem is convex), we know that the equalities will hold at the saddle point.

Given the saddle point $(\mu^\star, w^\star, \xi^\star)$, to derive $w^\star, \delta^\star$ from $\mu^\star$, we know that $w^\star, \delta^\star = \arg\min_{w,\delta} \tilde{P}_\rho(w,\delta,\xi^\star)$. The result follows after computing $\nabla_w \tilde{P}_\rho(w,\delta,\xi^\star) = 0$ and $\nabla_\delta \tilde{P}_\rho(w,\delta,\xi^\star) = 0$. □

So our strategy for CRF learning is $\min_\xi \max_\mu D_\rho(\mu,\xi)$, since we know that

$$D_\rho(\mu^\star,\xi^\star) \equiv L(w^\star,\delta^\star,\mu^\star) \equiv P(w^\star).$$

Since we work on the space of μ and ξ, to compute the primal objectives or the duality gap, we can use the mapping specified by the optimality condition (5). More precisely, we define

$$w(\mu^{t,s}) = -\frac{1}{\lambda}\Psi\mu^{t,s}, \quad \delta(\mu^{t,s},\xi^t) = \xi^t - \frac{1}{\rho}A\mu^{t,s},$$

which is equivalent to the representer theorem. The above condition is also useful to recover intermediate $w^{t,s}$ from $\mu^{t,s}$, which allows us to test on the validation set or decide if we should stop the learning earlier.

B.5 Comparison with State-of-the-Art Structured Learning Methods

A number of recent works for CRF learning can be viewed as optimizing formulations which are exactly or fairly close to one of $P(w), D(\mu), P_\rho(w,\delta), \tilde{P}_\rho(w,\delta,\xi)$ or $D_\rho(\mu,\xi)$. In the following table, we compare these approaches, in terms of the optimization formulation, the convergence rate (respectively in the primal or in the dual), and the inference oracle used for computing the gradients (or blockwise gradients).

Table 1: The Comparison of Structured Learning Methods

Method	Learning Regime	Primal/Dual	Convergence	Inference Oracle
Meshi et al. (2010)	SSVM	Primal (w, δ)	Sublinear	graphwise MAP (inexact)
Hazan and Urtasun (2010)	LossAugCRF	Primal (w, δ)	Sublinear	graphwise marginal (inexact)
Lacoste-Julien et al. (2013)	SSVM	Dual (μ)	Sublinear	graphwise MAP
Schmidt et al. (2015)	CRF	Primal (w)	Linear	graphwise marginal
Tang et al. (2016)	CRF	Dual (μ)	Sublinear	graphwise MAP
Meshi et al. (2015)	SSVM (soft)	Dual $(\mu, \xi = 0)$	Sublinear	cliquewise MAP
Yen et al. (2016)	SSVM	Dual (μ, ξ)	Linear	cliquewise MAP
IDAL	LossAugCRF	Dual (μ, ξ)	Linear	cliquewise marginal

C Gini Oriented Tree-Reweighted Entropy

The Bethe entropy (Yedidia et al., 2005) is generally non-concave. Its concave counterparts, such as the tree-reweighted entropy (Wainwright et al., 2005) or the region-based entropy (London et al., 2015; Yedidia et al., 2005), are only concave on the local consistency polytope, but non-concave on $\mathcal{I} \backslash \mathcal{L}$ (i.e., when $A\mu \neq 0$). Indeed, the Bethe entropy and its concave variants are of the form $H_{\text{Bethe}}(\mu) = \sum_{i \in \mathcal{V}} c_i H_i(\mu_i) + \sum_{\{i,j\} \in \mathcal{E}} c_{ij} H_{ij}(\mu_{ij})$, where c_i and c_{ij} are counting numbers. Even when H_{Bethe} is concave on $\mathcal{L}$, some of the c_i or c_{ij} can be negative.

The construction of the oriented tree-reweighted entropy stems from the expression of the entropy of a directed tree as the sum of of the entropy of the root and the conditional entropies of the variable at each node given their parent variable. Precisely, for an oriented tree T with the root i_0, the joint entropy can be computed as

$$H_T(Y) := H(Y_{i_0}) + \sum_{j \to i \in T} H(Y_i \mid Y_j). \tag{6}$$

On a general graph, if T is a (directed) spanning tree of the graph, then

$$H_T(Y) := H(Y_{i_0}) + \sum_{t \to i \in T} H(Y_i \mid Y_j) \geq H(Y_{i_0}) + \sum_{k=1}^{m} H(Y_{i_k} \mid Y_{i_{k-1}}, \ldots, Y_{i_0}) =: H_{\text{Shannon}}(Y). \tag{7}$$

Thus, for any probability distribution over the set of valid directed spanning trees, in which tree T has probability ρ_T, the inequality above entails that $H_{\text{Shannon}}(Y) \leq \sum_T \rho_T H_T(Y) =: H_{\text{OTRW}}(Y)$, where $\rho_T \geq 0$ and $\sum_T \rho_T = 1$.

$H_T(Y)$ is concave since it is a sum of concave functions, and so is $H_{\text{OTRW}}(Y)$ (who is a convex combination of $H_T(Y)$). To see that, we need to prove the following fact.

Fact 1 (Concavity of the conditional entropy)**.** *The conditional entropy $H(Y_j \mid Y_i)$ is in fact a function of μ_{ij}, namely $H(Y_j \mid Y_i) = H(\mu_{ij}) - H(A_i \mu_{ij})$. Moreover, $H(Y_j \mid Y_i)$ is a concave function of μ_{ij}.*

Proof. By definition,

$$H(Y_j \mid Y_i) = \sum_{y_j, y_i} \mu_{ij}(y_j, y_i) \log \frac{\sum_{y_j} \mu_{ij}(y_j, y_i)}{\mu_{ij}(y_j, y_i)} = H(\mu_{ij}) - H(A_i \mu_{ij}).$$

To show $H(Y_j \mid Y_i)$ is concave, we compute its Hessian:

$$\begin{aligned} \frac{\partial^2 H(Y_j \mid Y_i)}{\partial \mu_{ij}^2} &= -\text{diag}\Big(\mathbf{1} \oslash \mu_{ij}\Big) + A^\intercal \text{diag}\Big(\mathbf{1} \oslash A\mu\Big) A \\ &= -\text{diag}\Big(\mathbf{1} \oslash \mu_{ij}\Big) + \text{diag}\Big(\{\frac{1}{\tilde{\mu}_i(y_i)} \mathbf{1}\mathbf{1}^\intercal\}_{y_i=1}^{k_i}\Big) \end{aligned}$$

where $\tilde{\mu}_i = A_i \mu_{ij}$, and $\oslash$ denotes entrywise division. Let's focus on the i-th block of the negative Hessian. To show that the i-th block is positive semidefinite, that is, that

$$\operatorname{diag}\Big(\Big\{\frac{1}{\mu_{ij}(y_i, y_j)}\Big\}_{1 \le y_j \le k_j}\Big) - \frac{1}{\tilde{\mu}_i(y_i)} \mathbf{1}\mathbf{1}^\intercal \succeq 0, \tag{8}$$

we can use the Schur complement condition for positive semidefiniteness. Let $U = \tilde{\mu}_i(y_i)$. Since $\tilde{\mu}_i(y_i) \succ 0$,

$$L - B^\intercal U^{-1} B \succeq 0 \quad \text{iff} \quad \begin{bmatrix} U & B \\ B^\intercal & L \end{bmatrix} = \begin{bmatrix} \tilde{\mu}_i(y_i) & \mathbf{1}^\intercal \\ \mathbf{1} & \operatorname{diag}\Big(\{\frac{1}{\mu_{ij}(y_i, y_j)}\}_{1 \le y_j \le k_j}\Big) \end{bmatrix} \succeq 0.$$

We also have $L = \operatorname{diag}\Big(\Big\{\frac{1}{\mu_{ij}(y_i, y_j)}\Big\}_{y_j=1}^{k_j}\Big) \succ 0$, then

$$\begin{bmatrix} U & B \\ B^\intercal & L \end{bmatrix} \succeq 0 \quad \text{iff} \quad U - BL^{-1}B^\intercal = \tilde{\mu}_i(y_i) - \mathbf{1}^\intercal \operatorname{diag}\Big(\{\mu_{ij}(y_i, y_j)\}_{y_j=1}^{k_j}\Big)\mathbf{1} = \tilde{\mu}_i(y_i) - \tilde{\mu}_i(y_i) \succeq 0.$$

Because the last inequality holds, we know (8) must be true, which implies that the Hessian of $H(Y_j \mid Y_i)$ is negative semidefinite, thus $H(Y_j \mid Y_i)$ is concave. □

Note that $H_{\mathrm{OTRW}}(\mu)$ is concave on the entire set $\mathcal{I}$, unlike many Bethe entropy variants who are only concave in the local consistency polytope.

We define $\overrightarrow{\mathcal{E}}$ the directed edge set by expanding each edge from $\mathcal{E}$ with two directed edges, ρ_i and $\rho_{i|j}$ respectively as the probabilities of i (as the root) and $i \to t$ appearing in an oriented spanning tree when the latter is drawn with probability ρ_T. Then the oriented tree-reweighted entropy takes the form

$$\begin{aligned} H_{\mathrm{OTRW}}(\mu) := & \sum_{\{i,j\} \in \mathcal{E}} \rho_{j|i}\big[H_e(\mu_{ij}) - H_i(A_i \mu_{ij})\big] \\ & + \rho_{i|j}\big[H_e(\mu_{ij}) - H_j(A_j \mu_{ij})\big] + \sum_{i \in \mathcal{V}} \rho_i H_i(\mu_i), \end{aligned} \tag{9}$$

where $H_i(\mu_i) = -\sum_{y_i} \mu_i(y_i) \log \mu_i(y_i)$, $H_e(\mu_{ij}) = -\sum_{y_i, y_j} \mu_i(y_i, y_j) \log \mu_i(y_i, y_j)$ and $\rho_i, \rho_{i|j}, \rho_{j|i}$ are node/edge appearance probabilities in $[0, 1]$. H_{OTRW} is concave, since H_i is concave and it can be checked that so is $\mu_{ij} \mapsto H_e(\mu_{ij}) - H_i(A_i \mu_{ij})$ (although not strongly concave). It is easy to precompute the appearance probabilities ρ_i and $\rho_{i|j}$ via a variant of the directed matrix-tree theorem. See Koo et al. (2007) for more details.

A generic difficulty with entropies, is that H_i and H_e do not have Lipschitz gradients, which prevents the direct application of proximal methods with usual quadratic proximity terms. We thus propose to replace H_i and H_e by their second-order Taylor approximation around the uniform distribution. This yields a surrogate of the form

$$\begin{aligned} H_{\mathrm{GTRW}}(\mu) := & \sum_{\{i,j\} \in \mathcal{E}} \varepsilon\Big[k_i \rho_{j|i} \|A_i \mu_{ij}\|^2 + k_j \rho_{i|j} \|A_j \mu_{ij}\|^2\Big] \\ & - k_i k_j (\rho_{i|j} + \rho_{j|i}) \|\mu_{ij}\|^2 + \sum_{i \in \mathcal{V}} k_i \rho_i (1 - \|\mu_i\|^2), \end{aligned} \tag{10}$$

where $\varepsilon = 1$. Since this function is not strongly convex w.r.t. μ_{ij} because $k_j I_{k_i} - A_i^\intercal A_i$ has a non-trivial kernel, so we also consider variants with $\varepsilon < 1$ and denote them $H_{\mathrm{GTRW},\varepsilon}$. We call this approximation the *Gini OTRW entropy*, since it is consistent with the definition of Gini conditional entropy of Furuichi (2006).

D Proof of Lemma 1 and associated lemma

To prove Lemma 1, we first need to show $d(\xi)$ is a smooth function, and then we build up the associated lemmas which will be used in the proof of Lemma 1. Finally, in the end of this section, we prove Corollary 2 as a result to show the linear convergence in the primal.

D.1 Smoothness of $d(\xi)$

Lemma D.1. *(Hong and Luo, 2017, Lemma 2.3) $d(\xi)$ is convex and L_d-smooth, where $L_d \leq \rho$.*

Proof. By definition we have

$$D_\rho(\mu, 0) = -\langle \mu, \ell \rangle + \gamma F_{\mathcal{I}}^*(\mu) + \frac{1}{2\lambda}\|\Psi\mu\|^2 + \frac{1}{2\rho}\|A\mu\|^2.$$

We then have $d(\xi) = \max_\mu D_\rho(\mu, \xi) = \max_\mu \langle \mu, A^\intercal \xi \rangle - D_\rho(\mu, 0)$ so that if $J(\mu) := D_\rho(\mu, 0)$, then $d(\xi) = J^*(A^\intercal \xi)$ and d is a convex function by Fenchel conjugacy.

For any ξ_1 and ξ_2, denote by μ_1 and μ_2 the minimizers of $D_\rho(\cdot, \xi_1)$ and $D_\rho(\cdot, \xi_2)$ respectively. By convexity of $d(\xi)$ and the definition of subgradient, there exists $s_1 \in \partial F_{\mathcal{I}}^*(\mu_1)$ and $s_2 \in \partial F_{\mathcal{I}}^*(\mu_2)$ such that

$$A^\intercal \xi_1 + \ell - \gamma s_1 - \frac{1}{\lambda}\Psi^\intercal \Psi \mu_1 - \frac{1}{\rho} A^\intercal A \mu_1 = 0$$
$$A^\intercal \xi_2 + \ell - \gamma s_2 - \frac{1}{\lambda}\Psi^\intercal \Psi \mu_2 - \frac{1}{\rho} A^\intercal A \mu_2 = 0$$

By convexity of $F_{\mathcal{I}}^*(\mu)$, we have

$$\langle s_1 - s_2, \mu_1 - \mu_2 \rangle \geq 0,$$

which together with the equations above yields

$$\langle A^\intercal(\xi_1 - \xi_2) - \frac{1}{\lambda}\Psi^\intercal \Psi(\mu_1 - \mu_2) - \frac{1}{\rho} A^\intercal A(\mu_1 - \mu_2), \mu_1 - \mu_2 \rangle \geq 0.$$

Hence,

$$\langle \xi_1 - \xi_2, A(\mu_1 - \mu_2) \rangle \geq \frac{1}{\lambda}\|\Psi(\mu_1 - \mu_2)\|^2 + \frac{1}{\rho}\|A(\mu_1 - \mu_2)\|^2 \geq \frac{1}{\rho}\|A(\mu_1 - \mu_2)\|^2.$$

Now substituting $\nabla d(\xi_1) - \nabla d(\xi_2) = A(\mu_1 - \mu_2)$ into the above inequality and using the Cauchy-Schwarz inequality yields

$$\|\nabla d(\xi_1) - \nabla d(\xi_2)\| \leq \rho \|\xi_1 - \xi_2\|.$$

That completes the proof. □

D.2 Associated lemmas for Lemma 1

We first quantify in the next two lemmas how much $D(\mu, \xi^t)$ should be minimized in μ to provide a sufficiently accurate approximate gradient that it guarantees descent on d.

Lemma D.2 (Error on the gradient)**.** *Denote $\bar{\mu}^t := \mu^\star(\xi^t) = \operatorname{argmin}_\mu D(\mu, \xi^t)$; $g_t := \nabla d(\xi^t) = A\mu^\star(\xi^t)$ and $\hat{g}_t := A\hat{\mu}^t$. Let $\hat{\Delta}_t := D_\rho(\bar{\mu}^t, \xi^t) - D_\rho(\hat{\mu}^t, \xi^t)$. We have $\frac{1}{2L_d}\|\hat{g}_t - g_t\|^2 \leq \hat{\Delta}_t$, where L_d is the smoothness constant of d.*

Proof. Let $d^*(y) = \max_\xi \langle \xi, y \rangle - d(\xi)$. Then, it can easily be checked by using the definition of d and exchanging the order of maximization and minimization that $d^*(y) = \min_\mu D_\rho(\mu, 0) + \iota_{\{A\mu = y\}}$,

Since d is convex, we have $d(\xi) = \max_y \langle \xi, y \rangle - d^*(y)$, so that if $y^\star(\xi)$ is a maximizer for fixed ξ we have

$$0 \in \xi - \partial d^*(y^\star(\xi)) \Rightarrow \xi \in \partial d^*(y^\star(\xi)).$$

The strong convexity of $d^*(y)$ implies that, for all y,

$$d^*(y) - d^*(y^\star(\xi)) - \langle \xi, y - y^\star(\xi) \rangle \geq \frac{1}{2L_d}\|y - y^\star(\xi)\|^2.$$

But for any μ, we have $D_\rho(\mu,\xi) = \langle A\mu,\xi\rangle - D_\rho(\mu,0) \le \langle A\mu,\xi\rangle - d^*(A\mu)$, and, for $\mu^\star(\xi)$, this inequality is an equality, since we have $D_\rho(\mu^\star(\xi),\xi) = \langle y^\star(\xi),\xi\rangle - d^*(y^\star(\xi))$ and $y^\star(\xi) = A\mu^\star(\xi)$. As a consequence, setting $y = A\mu$, we have

$$D_\rho(\mu^\star(\xi),\xi) - D_\rho(\mu,\xi) \ge \frac{1}{2L_d}\|A\mu - A\mu^\star(\xi)\|^2$$

by definition of $D_\rho(\mu,\xi)$. We conclude the proof by substituting μ with $\hat{\mu}^t$ and ξ with ξ^t. □

Lemma D.3 (Guaranteed decrease on d)**.** *If we take inexact gradient on ξ with a fixed step size $\frac{1}{L_d}$, namely $\xi^{t+1} = \xi^t - \frac{1}{L_d}\hat{g}_t$, then*

$$d(\xi^t) - d(\xi^{t+1}) \ge \frac{\tau}{L_d}\Gamma_t - \hat{\Delta}_t, \tag{11}$$

where $\tau \in (0, L_d)$ satisfying $\frac{1}{2\tau}\|g_t\|^2 \ge \Gamma_t$.

Proof. Since $d(\xi)$ is L_d-smooth, we have

$$d(\xi^{t+1}) - d(\xi^t) \le \langle \nabla d(\xi^t), \xi^{t+1} - \xi^t\rangle + \frac{L_d}{2}\|\xi^{t+1} - \xi^t\|^2$$

Using the gradient step and $\nabla d(\xi^t) = g_t$, the above inequality can be simplified as

$$\begin{aligned} d(\xi^{t+1}) - d(\xi^t) &\le \langle g_t, -1/L_d\hat{g}_t\rangle + \frac{L_d}{2}\|1/L_d\hat{g}_t\|^2 \\ &= \frac{1}{2L_d}\Big(\|\hat{g}_t - g_t\|^2 - \|g_t\|^2\Big). \end{aligned} \tag{12}$$

We notice that the error bound given by the Lemma 2.3 of Hong and Luo (2017) holds for $d(\xi)$. Specifically,

$$\exists \tau' > 0, \text{ such that } \|\nabla d(\xi)\| \ge \tau'\|\xi - \xi^\star\|.$$

Since $d(\xi)$ is L_d-smooth and $\nabla d(\xi^\star) = 0$, we have

$$d(\xi) - d(\xi^\star) \le \frac{L_d}{2}\|\xi - \xi^\star\|^2 \le \frac{L_d}{2\tau'}\|\nabla d(\xi)\|^2,$$

which implies

$$\frac{1}{2\tau}\|g_t\|^2 \ge \Gamma_t,$$

where $\tau = \frac{\tau'}{L_d}$. By using (12) and the above inequality on $\|g_t\|^2$, we obtain

$$d(\xi^t) - d(\xi^{t+1}) \ge \frac{1}{2L_d}\Big(\|g_t\|^2 - \|\hat{g}_t - g_t\|^2\Big) \ge \frac{\tau}{L_d}\Gamma_t - \hat{\Delta}_t.$$

□

Since for each value of ξ^t the value and gradient of $d(\xi^t)$ need to be computed approximately by minimizing the augmented Lagrangian $D_\rho(\cdot,\xi^t)$, and since the difference between two consecutive strongly convex objectives is $D_\rho(\mu,\xi^t) - D_\rho(\mu,\xi^{t-1}) = \langle \xi^{t-1} - \xi^t, A\mu\rangle$, which is a function that converges to zero when if the sequence $\{\xi^t\}_t$ converges, a warm-restart strategy using $\hat{\mu}^t$ as the initial point to the subproblem $\max_\mu D_\rho(\mu,\xi^{t+1})$ is beneficial, as characterized by the following lemma.

Lemma D.4 (Dual gap at warm start)**.** *Denote $\Delta^0_{t+1} := D_\rho(\bar{\mu}^{t+1},\xi^{t+1}) - D_\rho(\mu^{t+1,0},\xi^{t+1})$. If we let $\mu^{t+1,0} = \hat{\mu}^t$, then*

$$\Delta^0_{t+1} \le (4 + \frac{2}{\omega})\hat{\Delta}_t + (1+2\omega)\Gamma_t, \quad \forall\omega > 0. \tag{13}$$

Proof. By definition, we have $D_\rho(\bar{\mu}^{t+1}, \xi^{t+1}) = D_\rho(\mu^{t+1,*}, \xi^{t+1}) = d(\xi^{t+1})$. The initial gap of μ at iteration t can then be decomposed as

$$\begin{aligned}
\Delta_{t+1}^0 &= D_\rho(\bar{\mu}^{t+1}, \xi^{t+1}) - D_\rho(\mu^{t+1,0}, \xi^{t+1}) + d(\xi^t) - d(\xi^t) - D_\rho(\hat{\mu}^t, \xi^t) + D_\rho(\hat{\mu}^t, \xi^t) \\
&= \Big[d(\xi^t) - D_\rho(\hat{\mu}^t, \xi^t)\Big] + \Big[D_\rho(\hat{\mu}^t, \xi^t) - D_\rho(\mu^{t+1,0}, \xi^{t+1})\Big] + D_\rho(\bar{\mu}^{t+1}, \xi^{t+1}) - d(\xi^t) \\
&= \Big[D_\rho(\bar{\mu}^t, \xi^t) - D_\rho(\hat{\mu}^t, \xi^t)\Big] + \Big[D_\rho(\hat{\mu}^t, \xi^t) - D_\rho(\hat{\mu}^t, \xi^{t+1})\Big] + d(\xi^{t+1}) - d(\xi^t) \\
&= \hat{\Delta}_t + \frac{1}{L_d}\|\hat{g}_t\|^2 + d(\xi^{t+1}) - d(\xi^t)
\end{aligned}$$

Again, we used the gradient step $\xi^{t+1} = \xi^t - \frac{1}{L_d}\hat{g}_t$, and recall that $A\hat{\mu}^t = \hat{g}_t$.

Now, we can bound the term $\|\hat{g}_t\|^2$ from above using the fact that

$$\begin{aligned}
\frac{1}{L_d}\|\hat{g}_t\|^2 &= \frac{1}{L_d}\Big[\|g_t\|^2 + 2\langle g_t, \hat{g}_t - g_t\rangle + \|\hat{g}_t - g_t\|^2\Big] \\
&\leq \frac{1}{L_d}\Big[(1+\omega)\|g_t\|^2 + (1+1/\omega)\|\hat{g}_t - g_t\|^2\Big],
\end{aligned}$$

where the last inequality stems from the Cauchy-Schwarz inequality $\langle g_t, \hat{g}_t - g_t\rangle \leq \|g_t\|\,\|\hat{g}_t - g_t\|$ and the fact that for any any $a, b \in \mathbb{R}$ and $\omega > 0$, we have $2ab \leq \omega a^2 + b^2/\omega$.

Combining the upper bound of $d(\xi^{t+1}) - d(\xi^t)$ from (12), we get

$$\Delta_{t+1}^0 \leq \hat{\Delta}_t + \frac{3\omega + 2}{2\omega L_d}\|\hat{g}_t - g_t\|^2 + \frac{2\omega + 1}{2L_d}\|g_t\|^2. \tag{14}$$

Here, we can use again Lemma D.2 and the fact that $\frac{1}{2L_d}\|g_t\|^2 \leq \Gamma_t$, which is due to the smoothness of $d(\xi)$. It follows that

$$\Delta_{t+1}^0 \leq \big(4 + \tfrac{2}{\omega}\big)\,\hat{\Delta}_t + \big(1 + 2\omega\big)\,\Gamma_t, \quad \forall \omega > 0.$$

□

D.3 Proof of Lemma 1

Combining Lemma D.3 and D.4, we now show that IDAL enjoys a linear convergence rate if we take a fixed number of inner iterations to estimate the gradient.

Lemma 1 (Linear convergence of the outer iteration)**.** *Suppose we have an algorithm $\mathcal{A}$ to approximately solve $\max_\mu D_\rho(\mu, \xi^t)$ in the sense that*

$$\exists \beta \in (0,1), \qquad \mathbb{E}[\hat{\Delta}_t] \leq \beta\,\mathbb{E}[\Delta_t^0].$$

Then $\exists \kappa \in (0,1)$ characterizing d and $C > 0$ such that, for any $\omega > 0$, after T_{ex} gradient steps on ξ, the suboptimalities $\Delta_{T_{\text{ex}}}$ and $\Gamma_{T_{\text{ex}}}$ are bounded from above:

$$\left\|\begin{matrix}\mathbb{E}[\hat{\Delta}_{T_{\text{ex}}}] \\ \mathbb{E}[\Gamma_{T_{\text{ex}}}]\end{matrix}\right\| \leq C\,\lambda_{\max}(\beta)^{T_{\text{ex}}} \left\|\begin{matrix}\mathbb{E}[\hat{\Delta}_0] \\ \mathbb{E}[\Gamma_0]\end{matrix}\right\|, \quad \text{where} \quad M(\beta) = \begin{bmatrix}\beta(4 + \frac{2}{\omega}) & \beta(1+2\omega) \\ 1 & 1 - \kappa\end{bmatrix}, \tag{15}$$

and $\lambda_{\max}(\beta)$ is the largest eigenvalue of $M(\beta)$. Thereby, if β is chosen so that $\lambda_{\max}(\beta) < 1$, Algorithm 1 is linearly convergent with a rate $\lambda_{\max}(\beta)$.

Proof. Note that $\Gamma_{t+1} - \Gamma_t = d(\xi^{t+1}) - d(\xi^t)$. By using Lemma D.3, we have an upper bound on Γ_{t+1} in terms of Γ_t and $\hat{\Delta}_t$, namely

$$\Gamma_{t+1} \leq \hat{\Delta}_t + (1 - \kappa)\,\Gamma_t \qquad \text{with} \qquad \kappa = \frac{\tau}{L_d}. \tag{16}$$

On the other hand, we can also derive an upper bound on $\hat{\Delta}_{t+1}$ in terms of Γ_t and $\hat{\Delta}_t$. To achieve that, we relate the inner problem with Γ_t by running the steps on μ until $\mathbb{E}[\hat{\Delta}_{t+1}] \leq (1-\pi)^{T_{\text{in}}}\mathbb{E}[\Delta^0_{t+1}] \leq \beta\,\mathbb{E}[\Delta^0_{t+1}]$, which means $T_{\text{in}} \geq \frac{\log\beta}{\log(1-\pi)}$. By Lemma D.4, we have

$$\mathbb{E}[\hat{\Delta}_{t+1}] \leq \beta\,\mathbb{E}[\Delta^0_{t+1}] \leq \beta\big(4+\tfrac{2}{\omega}\big)\mathbb{E}[\hat{\Delta}_t] + \beta(1+2\omega)\,\mathbb{E}[\Gamma_t]. \tag{17}$$

Combining (17) and (16), and taking expectations on both sides, we get

$$\begin{bmatrix}\mathbb{E}[\hat{\Delta}_{t+1}]\\ \mathbb{E}[\Gamma_{t+1}]\end{bmatrix} \leq M \begin{bmatrix}\mathbb{E}[\hat{\Delta}_t]\\ \mathbb{E}[\Gamma_t]\end{bmatrix} \tag{18}$$

Since by definition, all the elements of M are positive, we can telescope a sequence of matrix multiplications to get

$$\begin{bmatrix}\mathbb{E}[\hat{\Delta}_{T_{\text{ex}}}]\\ \mathbb{E}[\Gamma_{T_{\text{ex}}}]\end{bmatrix} \leq M \begin{bmatrix}\mathbb{E}[\hat{\Delta}_{T_{\text{ex}}-1}]\\ \mathbb{E}[\Gamma_{T_{\text{ex}}-1}]\end{bmatrix} \leq \cdots \leq M^{T_{\text{ex}}} \begin{bmatrix}\mathbb{E}[\hat{\Delta}_0]\\ \mathbb{E}[\Gamma_0]\end{bmatrix} \tag{19}$$

Assuming the eigen decomposition of M takes the form $M = PDP^{-1}$, then $M^t = PD^tP^{-1}$. Applying norms on both sides of the vector inequality, we have

$$\left\|\begin{bmatrix}\mathbb{E}[\hat{\Delta}_{T_{\text{ex}}}]\\ \mathbb{E}[\Gamma_{T_{\text{ex}}}]\end{bmatrix}\right\| \leq \|P\|_{\text{op}}\,\lambda_{\max}(\beta)^{T_{\text{ex}}}\|P^{-1}\|_{\text{op}} \left\|\begin{bmatrix}\mathbb{E}[\hat{\Delta}_0]\\ \mathbb{E}[\Gamma_0]\end{bmatrix}\right\|. \tag{20}$$

Note that $C = \|P\|_{\text{op}}\|P^{-1}\|_{\text{op}}$ is a constant. □

Corollary 2. *Let σ denote the strong convexity constant of $\mu \mapsto D_\rho(\mu,\xi)$ and L_d the smoothness constant of d. Assume that $(\|\xi_t\|_2)_{t\in\mathbb{N}}$ is almost surely bounded by a constant B. Then the squared residuals to the constraint $A\mu = 0$ satisfy*

$$\frac{1}{2}\|A\hat{\mu}^t\|_2^2 \leq 2L_d\Gamma_t + \frac{2}{\sigma}\|A\|_{\text{op}}^2\hat{\Delta}_t.$$

Furthermore, if we let $D_\infty(\mu) := \langle \ell, \mu\rangle - \gamma F^*_{\mathcal{I}}(\mu) - \frac{1}{2\lambda}\|\Psi\mu\|_2^2$, so that we have $D(\mu) = D_\infty(\mu) - \iota_{\{A\mu=0\}}$, then (given that $\mu^t \in \mathcal{I}$ throughout the algorithm) the gap between the smooth part of the objective in $\hat{\mu}^t$ and at the optimum can be bounded as follows

$$|D_\infty(\hat{\mu}^t) - D_\infty(\mu^\star)| \leq B\sqrt{2L_d\,\Gamma_t} + B\frac{\|A\|_{\text{op}}}{\sqrt{\sigma}}\sqrt{2\hat{\Delta}_t} + \Big(1+2\frac{L_d}{\rho}\Big)\Gamma_t + \Big(1+2\frac{\|A\|_{\text{op}}^2}{\rho\sigma}\Big)\hat{\Delta}_t.$$

Finally, if Γ_t and $\hat{\Delta}_t$ converge to 0 linearly then both the residuals $\|A\hat{\mu}^t\|_2^2$ and the gap in objective value $|D_\infty(\hat{\mu}^t) - D_\infty(\mu^\star)|$ converge to 0 linearly.

Proof. For the first inequality, by Fact D.1 we know that d is an L_d-smooth function. It is then a standard result (see e.g. Nesterov, 2013, Thm 2.1.5) that we therefore have $\|\nabla d(\xi^t)\|_2^2 \leq 2L_d(d(\xi^t) - d(\xi^\star)) = 2L_d\Gamma_t$. But since $\nabla d(\xi^t) = A\bar{\mu}^t$, and using the strong convexity of $\mu \mapsto D_\rho(\mu,\xi)$, we have

$$\frac{1}{2}\|A\hat{\mu}^t\|_2^2 \leq \|A\bar{\mu}^t\|_2^2 + \|A\|_{\text{op}}^2\|\bar{\mu}^t - \hat{\mu}^t\|_2^2 \leq 2L_d\Gamma_t + 2\frac{\|A\|_{\text{op}}^2}{\sigma}\hat{\Delta}_t.$$

For the second inequality, by definition of D_∞, we have $D_\rho(\mu,\xi) := D_\infty(\mu) + \langle \xi, A\mu\rangle - \frac{1}{2\rho}\|A\mu\|_2^2$, and $D_\infty(\mu^\star) = D(\mu^\star)$.

But then

$$\begin{aligned}|D_\infty(\hat{\mu}^t) - D_\infty(\mu^\star)| &= |D_\infty(\hat{\mu}^t) - D_\rho(\hat{\mu}^t,\xi^t)| + |D_\rho(\hat{\mu}^t,\xi^t) - D_\rho(\bar{\mu}^t,\xi^t)| + |D_\rho(\bar{\mu}^t,\xi^t) - D_\infty(\mu^\star)|\\ &\leq |\langle \xi^t, A\hat{\mu}^t\rangle| + \frac{1}{2\rho}\|A\hat{\mu}^t\|_2^2 + \hat{\Delta}_t + \Gamma_t.\end{aligned}$$

but we then have $|\langle \xi^t, A\hat{\mu}^t\rangle| + \frac{1}{2\rho}\|A\hat{\mu}^t\|_2^2 \leq B\|A\bar{\mu}^t\| + B\|A\|_{\mathrm{op}}\|\hat{\mu}^t - \bar{\mu}^t\|_2 + \frac{1}{\rho}(\|A\hat{\mu}^t\|_2^2 + \|A\|_{\mathrm{op}}^2\|\hat{\mu}^t - \bar{\mu}^t\|_2^2)$, which yields the result using the same inequalities as before.

Finally, to show the implications of linear convergence, by Lemma 2.3 of Hong and Luo (2017), there exists $\tau' > 0$ such that $\|\nabla d(\xi)\| \geq \tau'\|\xi - \xi^\star\|$. So that, since $\|\nabla d(\xi^t)\|_2^2 \leq 2L_d\Gamma_t$, we have that if the sequence $(\Gamma_t)_{t\in\mathbb{N}}$ is bounded then so is $(\xi^t)_{t\in\mathbb{N}}$. Letting B be a bound on $\|\xi^t\|$ the previous statements shows the results. □

D.4 Proofs of Corollaries 3 and 4 (Total number of iterations)

Corollary 3. *To ensure that $\mathbb{E}\hat{\Delta}_t \leq \epsilon$ and $\mathbb{E}\hat{\Gamma}_t \leq \epsilon$ it is enough to run the algorithm for a total number of inner iteration $T_{\mathrm{tot}} := T_{\mathrm{in}}T_{\mathrm{ex}}$ such that*

$$T_{\mathrm{tot}} \geq \frac{\log(\beta)}{\log\lambda_{\max}(\beta)\log(1-\pi)}\log(\epsilon)$$

Proof. To guarantee that $(1-\pi)_{\mathrm{in}}^T < \beta$ requires that $T_{\mathrm{in}} \geq \frac{\log(1-\pi)}{\log(\beta)}$ and to guaranteed that $\lambda_{\max}(\beta)^{T_{\mathrm{ex}}} < \epsilon$ requires similarly that $T_{\mathrm{ex}} \geq \frac{\log(\epsilon)}{\log(\lambda(\beta))}$. Taking the product of these inequalities yields the result. □

Corollary 4. *Let $\Delta^\star_{t\,T_{\mathrm{in}}+s} := \Delta_t^s + \Gamma_t$. If $\kappa < \frac{1}{2}$ and $\alpha = \frac{1}{12}$, if $T_{\mathrm{in}} \geq \frac{\log(\alpha\kappa)}{\log(1-\pi)}$, then, there exist a constant $C' > 0$ such that after a total of i clique updates, we have*

$$\mathbb{E}[\Delta_s^\star] \leq C'\Big(1 - \frac{\kappa\pi}{2\log(12/\kappa)}\Big)^s.$$

Proof. Using solving the quadratic formula for the largest eigenvalue of a two-by-two matrix yields

$$\lambda_{\max}(\beta) = (1 - \kappa + 6\beta) + \sqrt{(1-\kappa-6\beta)^2 + 12\beta}.$$

It is immediate to verify that $\lambda_{\max}(\beta) < 1$ if and only if $\beta < \frac{1}{3}\frac{\kappa}{1+2\kappa}$. This shows that we need to choose $\beta = \alpha\kappa$ with $\alpha < \frac{1}{3(1+2\kappa)}$. So in particular, if $\alpha < \frac{1}{9}$, then the previous inequality is satisfied for any $0 < \kappa < 1$.

Moreover, if $\kappa \leq \frac{1}{2}$ and $\alpha < \frac{1}{6}$, we have $\lambda_{\max}(\beta) = \lambda_{\max}(\alpha\kappa) < 1 - \kappa(1-6\alpha)$. Indeed, letting $x = 3\beta$, and $\alpha' = 3\alpha$, we have

$$\begin{aligned}
2\lambda_{\max}(\beta) &= (1-\kappa+2x) + \sqrt{(1-\kappa-2x)^2 + 4x} \\
&= 1 - \kappa + 2x + \sqrt{(1-\kappa)^2 + 4x\kappa + 4x^2} \\
&= 1 - \kappa + 2\alpha'\kappa + \sqrt{(1-\kappa)^2 + 4\alpha'\kappa^2 + 4\alpha'^2\kappa^2} \\
&\leq 1 - \kappa + 2\alpha'\kappa + \sqrt{(1-\kappa)^2 + 4\alpha'\kappa(1-\kappa) + 4\alpha'^2\kappa^2} \\
&\leq 2\,(1 - \kappa + 6\alpha\kappa).
\end{aligned}$$

Setting $\alpha = \frac{1}{12}$, given that the rate r is $r = 1 - \exp\big(\frac{\log(1-\pi)\log(\lambda_{\max}(\beta))}{\log(\beta)}\big)$, we have

$$r \geq 1 - (1-\pi)^{\frac{\log(1-\frac{\kappa}{2})}{\log(\frac{\kappa}{12})}} \geq \frac{\log(1-\frac{\kappa}{2})}{\log(\frac{\kappa}{12})}\,\pi \geq \frac{\kappa}{-2\log(\frac{\kappa}{12})}\,\pi,$$

where, for the second and the third inequality, we used the fact that $\log(1-z) \geq z$ respectively for $z = \pi$ and for $z = -\frac{\kappa}{2}$.

□

E Details of Algorithm $\mathcal{A}$ and convergence proofs for SDCA

In this section, we specify the detailed form of $D_\rho(\mu, \xi)$, and show how to apply the proof scheme of Shalev-Shwartz and Zhang (2016) to SDCA for the maximization of $D_\rho(\mu, \xi)$ w.r.t. μ in order to prove Proposition 1. We first write a fully decomposed expression of $D_\rho(\mu, \xi)$. We have:

$$D_\rho(\mu,\xi) = \sum_{c\in\mathcal{C}} \langle \ell_c, \mu_c\rangle - f_c^*(\mu_c) - \frac{1}{2\lambda}\sum_{\tau\in\mathcal{T}} \Big\| \sum_{c\in\mathcal{C}_\tau} \Psi_c \mu_c \Big\|^2 - \frac{1}{2\rho}\sum_{\substack{e\in\mathcal{E}\\ i\in e}} \|\mu_i - A_i\mu_e\|^2 + \sum_{\substack{e\in\mathcal{E}\\ i\in e}} \langle \xi_{ei}, \mu_i - A_i\mu_e\rangle, \tag{21}$$

where $-f_c^*(\mu_c) = \gamma h_c(\mu_c) - \iota_{\triangle_c}(\mu_c)$.

We assume here that the entropy surrogate used is such that h_c is σ_c-strongly concave w.r.t. μ_c.

In particular this corresponds to two possible choices:

- The naive Gini entropy, for which $h_c(\mu_c) = (1 - \|\mu_c\|_2^2)$.
- The Gini-OTRW entropy (see Appendix C) for which, given positive numbers $\rho_i, \rho_{i|j}$ and $\rho_{j|i}$ for all nodes and edges, we have
  - $h_i(\mu_i) = \rho_i k_i (1 - \|\mu_i\|_2^2) \quad$ for $i \in \mathcal{V}$
  - $h_{ij}(\mu_{ij}) = h_{i|j}(\mu_{ij}) + h_{j|i}(\mu_{ij}) \quad$ for $\{i,j\} \in \mathcal{E} \quad$ with $h_{i|j}(\mu_{ij}) = k_i\rho_{i|j}(\varepsilon\|A_j\mu_{ij}\|_2^2 - k_j\|\mu_{ij}\|_2^2)$

  for $\varepsilon < 1$ which is σ_c-strongly concave in μ_c with $\sigma_i = 2k_i\rho_i$ if $i \in \mathcal{V}$ else $\sigma_{\{i,j\}} = 2(1-\varepsilon)k_ik_j(\rho_{i|j} + \rho_{j|i})$. (For $\varepsilon = 1$, the surrogate is not strongly concave, and a modification of the decomposition into a separable terms and a smooth term must be used to leverage strong convexity: see the discussion in Section 6.2 after Proposition 2).

The proof of convergence for SDCA is based on showing that the expected increase in dual objective provides an upper bound on a measure of duality gap. For the problem, we are considering the gap of interest is $\mathrm{gap}(w,\delta,\mu,\xi) := \tilde{P}_\rho(w,\delta,\xi) - D_\rho(\mu,\xi)$, which is an upper bound on the duality gap $P_\rho(w,\xi) - D_\rho(\mu,\xi)$. It can be decomposed as follows:

$$\begin{aligned}
\mathrm{gap}(w,\delta,\mu,\xi) = \quad & \Big[\gamma F_{\mathcal{I}}(\ell + \Psi^\intercal w + A^\intercal\delta) + \gamma F_{\mathcal{I}}^*(\mu) - \langle \ell + \Psi^\intercal w + A^\intercal\delta, \mu\rangle\Big] \\
& + \Big[\frac{\lambda}{2}\|w\|^2 + \frac{1}{2\lambda}\|\Psi\mu\|^2 - \langle -w, \Psi\mu\rangle\Big] + \Big[\frac{\rho}{2}\|\xi-\delta\|^2 + \frac{1}{2\rho}\|A\mu\|^2 - \langle \xi-\delta, A\mu\rangle\Big] \qquad (22)\\
= \quad & \Big[\sum_{c\in\mathcal{C}} f_c^*\big(\tfrac{1}{\gamma}\tilde{\theta}_c(w,\delta)\big) + f_c^*(\mu_c) - \langle \tilde{\theta}_c(w,\delta), \mu_c\rangle\Big] \\
& + \Big[\sum_{\tau\in\mathcal{T}} \frac{\lambda}{2}\|w_\tau\|^2 + \frac{1}{2\lambda}\|\sum_{c\in\mathcal{C}_\tau}\Psi_c\mu_c\|^2 - \langle -w_\tau, \sum_{c\in\mathcal{C}_\tau}\Psi_c\mu_c\rangle\Big] \\
& + \Big[\sum_{e\in\mathcal{E}}\sum_{i\in e} \frac{\rho}{2}\|\xi_{ei}-\delta_{ei}\|^2 + \frac{1}{2\rho}\|\mu_i - A_i\mu_i\|^2 - \langle \xi_{ei}-\delta_{ei}, \mu_i - A_i\mu_i\rangle\Big], \qquad (23)
\end{aligned}$$

where $\tilde{\theta}_c$ is defined by

$$\tilde{\theta}_c(w,\delta) := \begin{cases} \ell_i + \Psi_i^\intercal w_{\tau_i} + \sum_{e\ni s} \delta_{ei} & \text{for } c = i \in \mathcal{V}, \\ \ell_e + \Psi_e^\intercal w_{\tau_e} - \sum_{i\in e} A_i^\intercal \delta_{ei} & \text{for } c = e \in \mathcal{E}. \end{cases} \tag{24}$$

We now proceed to characterize the progress of the algorithm at each iteration, and to that end, we introduce appropriate notations. In particular, since ξ is fixed during the algorithm, we drop the dependance on ξ in different functions: Denote the objective of the subproblem w.r.t. clique c as

$$D_{\rho,c}(\mu_c, \mu^s_{-c}) := -f_c^*(\mu_c) - r(\mu_c, \mu^s_{-c}), \tag{25}$$

with r defined by

$$r(\mu_c, \mu^s_{-c}) := \frac{1}{2\lambda}\Big\| \sum_{b\in\mathcal{C}_{\tau_c}\setminus\{c\}} \Psi_b\mu^s_b + \Psi_c\mu_c\Big\|^2 + \begin{cases} \sum_{e\ni c} \frac{1}{2\rho}\|\mu_i - A_i\mu^s_e\|^2 - \langle \mu_i, \sum_{e\ni i}\xi_{ei} + \ell_i\rangle, & c = i \in \mathcal{V}, \\ \sum_{i\in e} \frac{1}{2\rho}\|\mu^s_i - A_i\mu_e\|^2 - \langle \mu_e, -\sum_{i\in e} A_i^\intercal \xi_{ei} + \ell_e\rangle, & c = e \in \mathcal{E}. \end{cases}$$

It is straightforward to show that r is convex and smooth with cliquewise smoothness constants

$$L_i = \frac{1}{\lambda}\mathrm{eig}_{\max}(\Psi_i^\intercal\Psi_i) + \frac{|\{e : e\ni i\}|}{\rho},\ i\in\mathcal{V} \quad \text{and} \quad L_e = \frac{1}{\lambda}\mathrm{eig}_{\max}(\Psi_e^\intercal\Psi_e) + \frac{1}{\rho}\sum_{i\in e}\mathrm{eig}_{\max}(A_i^\intercal A_i),\ e\in\mathcal{E}.$$

The proof of convergence hinges on the following key lemma.

Lemma E.1. *Taking one of the following updates on μ_c with μ_{-c} fixed:*

- $\mu_c^{s+1} = \arg\max_{\mu_c} D_{\rho,c}(\mu_c, \mu^s_{-c})$.
- *or, if $u \in \partial f_c(\tilde\theta_c(w^s, \delta^s))$, where f_c is the conjugate function of f_c^*.*

$$\textit{solve} \qquad \hat\alpha = \operatorname*{arg\,max}_{\alpha\in[0,1]} D_{\rho,c}(\mu_c^s + \alpha(u-\mu_c^s); \mu^s), \qquad \textit{and set} \qquad \mu_c^{s+1} = \mu_c^s + \hat\alpha(u-\mu_c^s).$$

Then, with $\pi = \min_{c\in\mathcal{C}} \frac{\sigma_c}{|\mathcal{C}|(\sigma_c+L_c)}$, the following inequality holds

$$\mathbb{E}_c[D_\rho(\mu^{s+1},\xi) - D_\rho(\mu^s,\xi)] \geq \pi\,\mathbb{E}_c[\tilde P_\rho(w^s,\delta^s,\xi) - D_\rho(\mu^s,\xi)], \quad \forall\xi,$$

where w^s, δ^s are updated to maintain the optimality conditions:

$$w^s = -\frac{1}{\lambda}\Psi\mu^s, \quad \delta^s = \xi - \frac{1}{\rho}A\mu^s.$$

Proof. Letting $\breve D_{\rho,c}$ be defined as,

$$\breve D_{\rho,c}(\mu_c;\mu^s) := -f_c^*(\mu_c) - r(\mu^s) - \langle \nabla_{\mu_c} r(\mu^s), \mu_c - \mu_c^s\rangle - \frac{L_c}{2}\|\mu_c - \mu_c^s\|^2,$$

we have $\breve D_{\rho,c}(\mu_c;\mu^s) \leq D_{\rho,c}(\mu_c)$, since $\mu_c \mapsto r(\mu_c, \mu^s_{-c})$ is L_c-smooth.

First, for the update $\mu_c^{s+1} = \arg\max_{\mu_c} D_{\rho,c}(\mu_c, \mu^s_{-c})$, we have that, for any direction $u - \mu_c^s$ and any step size $\alpha\in[0,1]$

$$\begin{aligned} D_\rho(\mu^{s+1},\xi) - D_\rho(\mu^s,\xi) &= D_{\rho,c}(\mu_c^{s+1}, \mu^s_{-c}) - D_{\rho,c}(\mu_c^s, \mu^s_{-c}) \\ &\geq D_{\rho,c}(\mu_c^s + \alpha(u-\mu_c^s), \mu^s_{-c}) - D_{\rho,c}(\mu_c^s, \mu^s_{-c}) \\ &\geq \breve D_{\rho,c}(\mu_c^s + \alpha(u-\mu_c^s); \mu^s) - D_{\rho,c}(\mu_c^s, \mu^s_{-c}). \end{aligned} \tag{26}$$

Showing the desired inequality for the second form of update thus implies the inequality for the first type of update. Expliciting $\breve D_{\rho,c}(\mu_c^s + \alpha(u-\mu_c^s);\mu^s)$, we have

$$\begin{aligned} \breve D_{\rho,c}(\mu_c^s + \alpha(u-\mu_c^s);\mu^s) = &- f_c^*(\mu_c^s + \alpha(u-\mu_c^s)) \\ &- r(\mu^s) - \langle \nabla_{\mu_c} r(\mu^s), \alpha(u-\mu_c^s)\rangle - \frac{\alpha^2 L_c}{2}\|u-\mu_c^s\|^2. \end{aligned} \tag{27}$$

Since $f_c^*(u)$ assumed σ_c-strongly convex, we have

$$f_c^*(\mu_c^s + \alpha(u-\mu_c^s)) \leq \alpha f_c^*(u) + (1-\alpha) f_c^*(\mu_c^s) - \frac{\sigma_c}{2}\alpha(1-\alpha)\|u-\mu_c^s\|_2^2. \tag{28}$$

Combining (27) and (28), we obtain

$$\breve{D}_{\rho,c}(\mu_c^s + \alpha(u-\mu_c^s);\mu^s) \geq -\alpha\Big(f_c^*(u) - f_c^*(\mu_c^s) + \langle \nabla_{\mu_c} r(\mu^s), u-\mu_c^s\rangle\Big) \\ - f_c^*(\mu_c^s) - r(\mu^s) + \Big(\frac{\sigma_c}{2}\alpha(1-\alpha) - \frac{\alpha^2 L_c}{2}\Big)\|u-\mu_c^s\|^2. \tag{29}$$

Now, if we choose $u \in \partial f_c\big(-\nabla_{\mu_c} r(\mu^s)\big)$, by Fenchel conjugacy, it follows that

$$f_c\big(-\nabla_{\mu_c} r(\mu^s)\big) = -f_c^*(u) - \langle \nabla_{\mu_c} r(\mu^s), u\rangle.$$

One can easily see that $\tilde{\theta}_c(w^s,\delta^s) = -\nabla_{\mu_c} r(\mu^s)$ by maintaining the optimality conditions

$$\forall c \in \mathcal{C}\colon \quad w_{\tau_c}^s = -\frac{1}{\lambda}\sum_{b\in\mathcal{C}_\tau} \Psi_b \mu_b^s, \qquad \forall e\in\mathcal{E}, i\in e\colon \quad \delta_{ei}^s = \xi_{ei} - \frac{1}{\rho}(\mu_i^s - A_i\mu_e^s).$$

Thus, we can further simplify (29) as

$$\breve{D}_{\rho,c}(\mu_c^s + \alpha(u-\mu_c^s);\mu^s) - D_{\rho,c}(\mu_c^s,\mu_{-c}^s) \geq \alpha\Big(f_c(\tilde{\theta}_c(w^s,\delta^s)) + f_c^*(\mu_c^s) - \langle \tilde{\theta}_c(w^s,\delta^s), \mu_c^s\rangle\Big), \tag{30}$$

provided that $\frac{\sigma_c}{2}\alpha(1-\alpha) - \frac{\alpha^2 L_c}{2} \geq 0$, that is, $0 \leq \alpha \leq \frac{\sigma_c}{\sigma_c + L_c}$.

The key observation is that

$$\mathrm{gap}(w,\delta,\mu,\xi) = \sum_{c\in\mathcal{C}} f_c(\tilde{\theta}_c(w,\delta)) + f_c^*(\mu_c) - \langle \tilde{\theta}_c(w,\delta), \mu_c\rangle \tag{31}$$

if we maintain the optimality conditions. By using (31) and taking expectation $\mathbb{E}_c$ w.r.t. a uniform random choice of the clique c on both sides of (30), we guarantee that, for $\alpha \in [0, \frac{\sigma_c}{\sigma_c+L_c}]$,

$$\mathbb{E}_c[D_\rho(\mu^{s+1},\xi) - D_\rho(\mu^s,\xi)] \geq \mathbb{E}_c\Big[\frac{\alpha}{|\mathcal{C}|}\mathrm{gap}(w^s,\delta^s,\mu^s,\xi)\Big].$$

So, we can choose the maximum value $\frac{\sigma_c}{\sigma_c+L_c}$ for α. It follows that

$$\mathbb{E}_c[D_\rho(\mu^{s+1},\xi) - D_\rho(\mu^s,\xi)] \geq \Big(\min_{c\in\mathcal{C}} \frac{\sigma_c}{|\mathcal{C}|(\sigma_c+L_c)}\Big)\,\mathbb{E}_c[\,\mathrm{gap}(w^s,\delta^s,\mu^s,\xi)].$$

□

We can now prove Proposition 1.

Proposition 1. *If $\mathcal{A}$ is SDCA, let $|\mathcal{C}|$ be the total number of cliques, σ_c the strong convexity constant of f_c^*, and L_c the Lipschitz constant of $\mu_c \mapsto r(\mu)$, then $\mathcal{A}$ is linearly convergent with rate $\pi = \min_{c\in\mathcal{C}} \frac{\sigma_c}{|\mathcal{C}|(\sigma_c+L_c)}$.*

Proof. Denote $\Delta_t^s := D_\rho(\bar{\mu}^t,\xi^t) - D_\rho(\mu^{t,s},\xi^t)$. Since we update $\mu^{t,s}$ to $\mu^{t,s+1}$ using SDCA, according to Lemma E.1, we have

$$\begin{aligned}\mathbb{E}_c[\Delta_t^s - \Delta_t^{s+1}] &= \mathbb{E}_c[D_\rho(\mu^{t,s+1},\xi^t) - D_\rho(\mu^{t,s},\xi^t)] \\ &\geq \pi\,\mathbb{E}_c[\tilde{P}_\rho(w(\mu^{t,s}),\delta(\mu^{t,s},\xi^t),\xi^t) - D_\rho(\mu^{t,s},\xi^t)] \\ &\geq \pi\,\mathbb{E}_c[D_\rho(\bar{\mu}^t,\xi^t) - D_\rho(\mu^{t,s},\xi^t)] = \pi\mathbb{E}_c[\Delta_t^s],\end{aligned}$$

and $\pi = \min_{c\in\mathcal{C}} \frac{\sigma_c}{|\mathcal{C}|(\sigma_c+L_c)}$. The above inequality implies that

$$\mathbb{E}_c[\Delta_t^{s+1}] \leq (1-\pi)\,\mathbb{E}_c[\Delta_t^s] \leq (1-\pi)^{s+1}\,\mathbb{E}_c[\Delta_t^0].$$

The result follows if we set $T_{\mathrm{in}} = s+1$.

□

E.1 Proof of Propositions 2 and 3 (Linear convergence in the primal)

Proposition 2. *Let $\hat{w}^t = w(\hat{\mu}^t)$. If $\mathcal{A}$ is SDCA, then*

$$\mathbb{E}[P(\hat{w}^t) - P(w^\star)] \leq \mathbb{E}[\frac{1}{\pi}\hat{\Delta}_t + \Gamma_t].$$

Proof. Recall that $P(w^\star) = D(\mu^\star) = D_\rho(\mu^\star, \xi^\star)$ by Corollary 5.

$$\begin{aligned}
P(w^{t,s}) - P(w^\star) &= P(w^{t,s}) - D_\rho(\bar{\mu}^t, \xi^t) + D_\rho(\bar{\mu}^t, \xi^t) - P(w^\star) \\
&= P(w^{t,s}) - D_\rho(\bar{\mu}^t, \xi^t) + D_\rho(\bar{\mu}^t, \xi^t) - D_\rho(\mu^\star, \xi^\star) \\
&\leq \tilde{P}(w^{t,s}, \delta^{t,s}, \xi^t) - D_\rho(\bar{\mu}^t, \xi^t) + d(\xi^t) - d(\xi^\star) \\
&\leq \tilde{P}(w^{t,s}, \delta^{t,s}, \xi^t) - D_\rho(\mu^{t,s}, \xi^t) + d(\xi^t) - d(\xi^\star) \\
&= \operatorname{gap}(w^{t,s}, \delta^{t,s}, \mu^{t,s}, \xi^t) + \Gamma_t
\end{aligned}$$

If $\mathcal{A}$ is SDCA, by Lemma E.1, we have

$$\begin{aligned}
\mathbb{E}[P(w^{t,s}) - P(w^\star)] &= \mathbb{E}[\operatorname{gap}(w^{t,s}, \delta^{t,s}, \mu^{t,s}, \xi^t) + \Gamma_t] \\
&\leq \mathbb{E}\Big[\frac{1}{\pi}(\Delta_t^s - \Delta_t^{s+1}) + \Gamma_t\Big] \\
&\leq \frac{1}{\pi}\mathbb{E}[\Delta_t^s] + \mathbb{E}[\Gamma_t].
\end{aligned}$$

Given that $\hat{\Delta}_t = \Delta_t^{T_{\text{in}}}$, the result follows by setting $s = T_{\text{in}}$. □

Proposition 3. *Let $w^{t,s} = w(\mu^{t,s})$. If $\mathcal{A}$ is a linearly convergent algorithm and $\mu \mapsto -H_{\text{approx}} + \frac{1}{2\rho}\|A\mu\|_2^2$ is strongly convex then $P(w^{t,s}) - P(w^\star)$ converges to 0 linearly.*

Proof. Note that, if σ is the strong convexity constant of D_ρ w.r.t. μ, then given that $P_\rho(w, \xi) = \min_\delta \tilde{P}_\rho(w, \delta, \xi)$ with

$$\tilde{P}_\rho(w, \delta, \xi) = \gamma F_{\mathcal{I}}\Big(\frac{\theta(w) + A^\intercal\delta}{\gamma}\Big) + \frac{\rho}{2}\|\delta - \xi\|^2 + \frac{\lambda}{2}\|w\|_2^2,$$

we also have

$$P_\rho(w, \xi) = \max_\mu \Big[\, \langle \mu, \Psi^\intercal w\rangle + \gamma H_{\text{approx}}(\mu) + \langle \xi, A\mu\rangle - \frac{1}{2\rho}\|A\mu\|_2^2\Big] + \frac{\lambda}{2}\|w\|_2^2,$$

which shows that $w \mapsto P_\rho(w, \xi)$ is a function with Lipschitz gradient as the sum of $w \mapsto \frac{\lambda}{2}\|w\|_2^2$ and of the Fenchel conjugate of a strongly convex function. Let L_P be its Lipschitz smoothness constant and note that it does not depend on the value of ξ. We thus have

$$P_\rho(w^{t,s}, \xi^t) - P_\rho(\bar{w}^t, \xi^t) \leq L_P\|w^{t,s} - \bar{w}^t\|_2^2.$$

Then given the representer theorem, and by strong convexity of $\mu \mapsto D_\rho(\mu, \xi)$ we have

$$\|w^{t,s} - \bar{w}^t\|_2^2 = \|\Psi(\mu^{t,s} - \bar{\mu}^t)\|_2^2 \leq \frac{1}{\sigma}\|\Psi\|_{\text{op}}^2(D_\rho(\mu^{t,s}, \xi^t) - D_\rho(\bar{\mu}^t, \xi^t))$$

So that, since $P(w^{t,s}) \leq P_\rho(w^{t,s}, \xi^t)$ and $P(w^\star) = P_\rho(w^\star, \xi^\star)$, we have

$$P(w^{t,s}) - P(w^\star) \leq P_\rho(w^{t,s}, \xi^t) - P_\rho(\bar{w}^t, \xi^t) + P_\rho(\bar{w}^t, \xi^t) - P_\rho(w^\star, \xi^\star) \leq \frac{L_P}{\sigma}\|\Psi\|_{\text{op}}^2\,\Delta_t^s + \Gamma_t.$$

Finally, global linear convergence in the primal also follows from the linear convergence of $\hat{\Delta}_t$ and Γ_t. □

F Notation summary

Given the number of notations in the main paper, we summarize some of them in Tables 2,3 and 4. The block matrices Ψ and A are schematically drawn below to illustrate their structure.

$$\Psi = \begin{array}{c} \\ \tau_c \end{array} \begin{array}{c} \begin{array}{c} c \end{array} \\ \begin{bmatrix} \cdots & \Psi_c \end{bmatrix} \end{array} \qquad A = \begin{array}{c} \\ ij \end{array} \begin{array}{c} \begin{array}{cc} i & ij \end{array} \\ \begin{bmatrix} \cdots & I_{k_i} & -A_i \end{bmatrix} \end{array}$$

Table 2: Notations for sets

Notation	Dimension	Description
$\mathcal{C}$		the set of cliques
$\mathcal{E}$		the set of edges
$\mathcal{V}$		the set of nodes
$\mathcal{Y}_i = \mathcal{S}_k$		$\mathcal{S}_k := \{u \in \{0,1\}^k \mid \|u\|_1 = 1\}$
$\mathcal{Y}_c$	$\prod_{i \in c} k_i$	$\mathcal{Y}_c := \bigtimes_{i \in c} \mathcal{Y}_i$
$\mathcal{Y}$	$\prod_{i \in V} k_i$	$\mathcal{Y} := \bigtimes_{i \in V} \mathcal{Y}_i$
$\mathcal{T}$		the set of clique types
$\mathcal{C}_\tau$		the set of cliques of type τ
$\mathcal{M}$		the marginal polytope
$\mathcal{L}$		the local polytope
$\mathcal{I}$		$\mathcal{I} := \prod_{i \in V} \Delta_{k_i} \times \prod_{e \in E} \Delta_{k_e}$

References

Furuichi, S. (2006). Information theoretical properties of Tsallis entropies. *Journal of Mathematical Physics*, 47(2):023302.

Hazan, T. and Urtasun, R. (2010). A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, pages 838–846.

Hong, M. and Luo, Z.-Q. (2017). On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199.

Koo, T., Globerson, A., Carreras Pérez, X., and Collins, M. (2007). Structured prediction models via the matrix-tree theorem. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 141–150.

Lacoste-Julien, S., Jaggi, M., Schmidt, M., and Pletscher, P. (2013). Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, pages 53–61.

London, B., Huang, B., and Getoor, L. (2015). The benefits of learning with strongly convex approximate inference. In *ICML*, pages 410–418.

Meshi, O., Sontag, D., Globerson, A., and Jaakkola, T. S. (2010). Learning efficiently with approximate inference via dual losses. In *ICML*, pages 783–790.

Meshi, O., Srebro, N., and Hazan, T. (2015). Efficient training of structured SVMs via soft constraints. In *AISTATS*, pages 699–707.

Nesterov, Y. (2013). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer.

Table 3: Notations for variables, parameters and functions

Notation	Domain	Description
k_i	$\mathbb{N}$	the number of values that Y_i can take
k_c	$\mathbb{N}$	$k_c := \lvert\mathcal{Y}_c\rvert = \prod_{i \in c} k_i$
τ	$\mathbb{N}$	the type of a clique
τ_c	$\mathbb{N}$	the type of clique c
w_τ	$\mathbb{R}^{d_\tau}$	the parameter shared by all cliques with type τ
w	$\mathbb{R}^{\sum_\tau d_\tau}$	$w := [w_\tau]_{\tau \in \mathcal{T}}$
$\phi_c(x, y_c)$	$\mathbb{R}^{d_{\tau_c}}$	the feature vector associated with the clique c given $Y_c = y_c$
$Z(x, w)$	$\mathbb{R}_{+*}$	the partition function of $p(y\|x; w)$
$\ell_c(y_c^{(n)}, y_c)$	$\mathbb{R}_+$	the user defined loss function associated with the clique c
γ	$(0, +\infty)$	the temperature parameter of the loss-augmented CRF
$\Psi_c^{(n)}$	$\mathbb{R}^{d_{\tau_c} \times k_c}$	$\Psi_c^{(n)} := \left[\phi_c(x^{(n)}, y_c) - \phi_c(x^{(n)}, y_c^{(n)})\right]_{y_c \in \mathcal{Y}_c}$
$\Psi^{(n)}$	$\mathbb{R}^{\sum_\tau d_\tau \times \sum_c k_c}$	see the drawing
$\ell_c^{(n)}$	$\mathbb{R}^{k_c}$	$\ell_c^{(n)} := \left[\ell_c(y_c^{(n)}, y_c)\right]_{y_c \in \mathcal{Y}_c}$
$\ell^{(n)}$	$\mathbb{R}^{\sum_c k_c}$	$\ell^{(n)} := [\ell_c^{(n)}]_{c \in C}$
$\theta_c^{(n)}(w)$	$\mathbb{R}^{k_c}$	$\theta_c^{(n)}(w) := {\Psi_c^{(n)}}^\intercal w_{\tau_c} + \ell_c^{(n)}$
$\theta^{(n)}(w)$	$\mathbb{R}^{\sum_c k_c}$	$\theta^{(n)}(w) := [\theta_c^{(n)}(w)]_{c \in \mathcal{C}}$, the natural parameter
F	$\mathbb{R}^{\sum_c k_c} \to \mathbb{R}$	the log partition function of θ
$T(y)$	$\mathbb{R}^{\sum_c k_c}$	the sufficient statistics
μ_c	$\mathbb{R}^{k_c}$	the mean parameter associated with the clique c
μ	$\mathbb{R}^{\sum_c k_c}$	the mean parameter
F^*	$\mathbb{R}^{\sum_c k_c} \to \mathbb{R}$	the Fenchel conjugate of F
ι_C	$\mathbb{R}^{\sum_c k_c} \to \{0, +\infty\}$	the indicator function of set C
λ	$\mathbb{R}_+$	the coefficient of the regularizer
A_i	$\mathbb{R}^{k_i \times k_e}$	the matrix encoding the marginalization constraint for i in e.
A	$\mathbb{R}^{\sum_e \sum_{i \in e} k_i \times \sum_c k_c}$	see the matrix form

Table 4: Notations smoothness, strong convexity constant and related quantities

Notation	Description
L_d	the Lipschitz constant of $\nabla d(\xi)$
τ	the constant of PL inequality for $d(\xi)$
σ_c	the strong convexity constant of $\mu_c \mapsto -H_{\text{Approx}}(\mu)$
L_c	the Lipschitz constant of $\mu_c \mapsto \frac{1}{\lambda}\Psi^\intercal w(\mu) + \frac{1}{\rho}A^\intercal \delta(\mu, \xi^t)$

Pletscher, P., Ong, C. S., and Buhmann, J. M. (2010). Entropy and margin maximization for structured output learning. In *ECML*, pages 83–98.

Schmidt, M., Babanezhad, R., Ahmed, M., Defazio, A., Clifton, A., and Sarkar, A. (2015). Non-uniform stochastic average gradient method for training conditional random fields. In *AIStats*, pages 819–828.

Shalev-Shwartz, S. and Zhang, T. (2016). Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1-2):105–145.

Tang, K., Ruozzi, N., Belanger, D., and Jebara, T. (2016). Bethe learning of graphical models via MAP decoding. In *AIStats*, pages 1096–1104.

Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. (2005). A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51(7):2313–2335.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312.

Yen, I. E.-H., Huang, X., Zhong, K., Zhang, R., Ravikumar, P. K., and Dhillon, I. S. (2016). Dual decomposed learning with factorwise oracle for structural SVM of large output domain. In *NIPS*, pages 5024–5032.