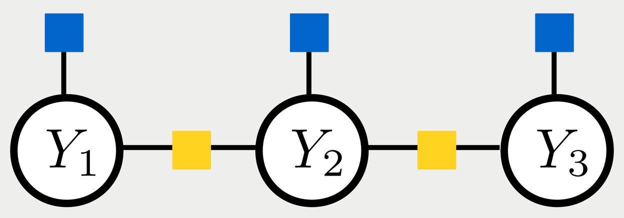
SDCA-Powered Inexact Dual Augmented Lagrangian Method for Fast CRF Learning

Shell Xu Hu (hus@imagine.enpc.fr) and Guillaume Obozinski (guillaume.obozinski@enpc.fr)
IMAGINE group, Laboratoire d'Informatique Gaspard Monge, École des Ponts ParisTech

1. Introduction

- ***Problem:*** *Maximum likelihood estimation of discrete conditional random fields with variational relaxation of the dual problem.*
- ***Method:*** *Dual augmented Lagrangian method with inexact inner-loop updates by SDCA.*

2. Conditional Random Fields



$\mathcal{T} = \{\tau \mid \tau = V \text{ or } E\}$, $\mathcal{C} = \mathcal{C}_V \cup \mathcal{C}_E$
$V = \{1, 2, 3\}$, $E = \{\{1, 2\}, \{1, 3\}\}$
$y_{12} = y_1 \otimes y_2$ (one-hot vectors)

- Given $\{(x^{(n)}, y^{(n)})\}_{1 \le n \le N}$, a CRF reads as

$$p(y^{(n)} \mid x^{(n)}; w) := \frac{1}{Z(x^{(n)}, w)} \prod_{\tau \in \mathcal{T}} \prod_{c \in \mathcal{C}_\tau} \exp\left(\langle w_\tau, \phi_c(x^{(n)}, y_c^{(n)})\rangle\right).$$

- Abstract CRF by using $\theta^{(n)}(w) := [\theta_c^{(n)}(w) := \Psi_c^{(n)\intercal} w_{\tau_c}]_{c \in \mathcal{C}} = \Psi^{(n)\intercal} w$ and $T(y) := [y_c]_{c \in \mathcal{C}}$ defined below:

$$\begin{aligned} -\log p(y^{(n)} \mid x^{(n)}; w) &= \log \sum_y \exp\left[\sum_{\tau \in \mathcal{T}} \sum_{c \in \mathcal{C}_\tau} \langle w_\tau, \phi_c(x^{(n)}, y_c) - \phi_c(x^{(n)}, y_c^{(n)})\rangle\right] \\ &= \log \sum_y \exp\left[\sum_{\tau \in \mathcal{T}} \sum_{c \in \mathcal{C}_\tau} \langle \Psi_c^{(n)\intercal} w_\tau, y_c\rangle\right] \\ &= \log \sum_y \exp\left[\langle \theta^{(n)}(w), T(y)\rangle\right] =: F\left(\theta^{(n)}(w)\right) \end{aligned}$$

3. Maximum Likelihood Estimation

- We assume $N = 1$: $\max_w \log p(y \mid x; w) \Leftrightarrow \min_w F\big(\theta(w)\big)$.
- *Computational issue:* $\nabla_{w_\tau} F(\theta(w)) = \sum_{c \in \mathcal{C}_\tau} \Psi_c \mathbb{E}_\theta[y_c]$ requires performing approximate marginal inference.

4. Variational Relaxation

- Fenchel conjugate form of F[4]:

$$F(\theta) = \max_\mu \left[\langle \mu, \theta\rangle - F^*(\mu)\right] \text{ with } F^*(\mu) = -H_{\text{Shannon}}(\mu) + \iota_{\mathcal{M}}(\mu).$$

where $\mathcal{M} := \{\mu \mid \mu = \mathbb{E}_\theta[T(Y)] \text{ for some } \theta\}$ is the marginal polytope.

- Relax $F \to F_{\mathcal{L}}$ by $\mathcal{M} \to \mathcal{L}$ and $H_{\text{Shannon}} \to H_{\text{Approx}}$:
  - $F_{\mathcal{L}}$ is defined similarly as F with $F_{\mathcal{L}}^*(\mu) := -H_{\text{Approx}}(\mu) + \iota_{\mathcal{I}}(\mu) + \iota_{A\mu=0}$.

$$\mathcal{L} := \overbrace{\{\mu \mid \forall c, i \in c\colon \mu_i(y_i) = \sum_{y_{c\setminus i}} \mu_c(y_c)\}}^{\equiv\{\mu \mid A\mu = 0\}} \bigcap \overbrace{\{\mu \mid \forall c\colon \mu_c \ge 0, \mu_c^\intercal \mathbf{1} = 1\}}^{=:\mathcal{I}}$$

  - H_{Approx} is block-separable, concave on $\mathcal{I}$ and strongly concave on $\mathcal{L}$.

5. "Inference-Free" Formulation

- The primal and dual of relaxed MLE:

$$\text{MLE}: \quad \min_w P(w) := F_{\mathcal{L}}\big(\theta(w)\big) + \frac{\lambda}{2}\|w\|_2^2$$

$$\text{MaxEnt}: \quad \max_\mu D(\mu) := -F_{\mathcal{L}}^*(\mu) - \frac{1}{2\lambda}\|\Psi\mu\|_2^2$$

- Augmented Lagrangian formulation for $A\mu = 0$ in the dual:

$$\min_\xi \max_\mu \Big[D_\rho(\mu, \xi) := \underbrace{H_{\text{Approx}}(\mu) - \iota_{\mathcal{I}}(\mu) + \langle \xi, A\mu\rangle}_{\text{block-separable \& concave}} \underbrace{- \frac{1}{2\rho}\|A\mu\|_2^2 - \frac{1}{2\lambda}\|\Psi\mu\|_2^2}_{\text{smooth}}\Big].$$

- For fixed ξ, it is natural to optimize $D_\rho(\mu, \xi)$ by stochastic coordinate ascent (e.g. SDCA[3]), so only clique-wise updates are needed.

References

[1] M. Hong and Z.-Q. Luo. "On the linear convergence of the alternating direction method of multipliers". In: *Mathematical Programming* 162.1-2 (2017), pp. 165–199.

[2] O. Meshi, N. Srebro, and T. Hazan. "Efficient Training of Structured SVMs via Soft Constraints". In: *AISTATS*. 2015, pp. 699–707.

[3] S. Shalev-Shwartz and T. Zhang. "Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization". In: *ICML*. 2014, pp. 64–72.

[4] M. J. Wainwright. "Graphical models, exponential families, and variational inference". In: *Foundations and Trends in Machine Learning* 1.1–2 (2008), pp. 1–305.

[5] I. H. Yen et al. "Dual Decomposed Learning with Factorwise Oracle for Structural SVM of Large Output Domain". In: *NIPS*. 2016, pp. 5024–5032.

6. Algorithm

- Optimization problem: $\min_\xi d(\xi)$ with $d(\xi) := \max_\mu D_\rho(\mu, \xi)$.
- $d(\xi)$ is L_d-smooth, τ-restricted-strongly-convex[1].
- IDAL: The idea is to solve $\min_\xi d(\xi)$ by an inexact gradient descent with warm restarts.

```
for t = 1, ..., T_ex:
    ξ^t = ξ^{t-1} − (1/L_d) A μ̂^{t-1};   μ^{t,0} = μ̂^{t-1}
    for s = 1, ..., T_in:
        Draw a clique c uniformly at random
        ν_c = prox_block_update(c, μ^{t,s-1})
        μ_c^{t,s} = ν_c;   μ_{-c}^{t,s} = μ_{-c}^{t,s-1};   μ̂^t = μ^{t,s} if s = T_in
```

- $\texttt{prox\_block\_update}(c, \mu)$ approximately $\max_{\mu_c} D_\rho([\mu_c, \mu_{-c}], \xi)$.

7. Analysis

Theorem 1 (Linear Convergence of the Outer Iteration)

- *Suboptimalities*: $\Gamma_t = d(\xi^t) - \min_\xi d(\xi)$, $\hat{\Delta}_t := \max_\mu D_\rho(\mu, \xi^t) - D_\rho(\hat{\mu}^t, \xi^t)$.
- SDCA on μ ensure $\mathbb{E}\hat{\Delta}_t \le (1 - \pi)^{T_{\text{in}}} \mathbb{E}\Delta_t^0$.
- If we run $T_{\text{in}} > \frac{\log(\beta)}{\log(1-\pi)}$ iterations on μ for $\beta \in (0, 1)$ with $\lambda_{\max}(\beta) < 1$, where $\lambda_{\max}(\beta)$ is the largest eigenvalue of $M(\beta)$ defined below, then after T_{ex} iterations on ξ we have

$$\left\|\begin{matrix}\mathbb{E}\hat{\Delta}_{T_{\text{ex}}} \\ \mathbb{E}\Gamma_{T_{\text{ex}}}\end{matrix}\right\| \le \text{const}\, \lambda_{\max}(\beta)^{T_{\text{ex}}} \left\|\begin{matrix}\mathbb{E}\hat{\Delta}_0 \\ \mathbb{E}\Gamma_0\end{matrix}\right\|, \quad \text{where } M(\beta) = \begin{bmatrix} 6\beta & 3\beta \\ 1 & 1 - \frac{\tau}{L_d}\end{bmatrix}.$$

Therefore, it is almost surely that $\hat{\Delta}_t, \Gamma_t$ converge linearly.

Corollary 1 (Bound on Total Inner Iterations)

To ensure that $\mathbb{E}\hat{\Delta}_t \le \epsilon$ and $\mathbb{E}\Gamma_t \le \epsilon$ it is enough to run $T_{\text{tot}} := T_{\text{in}} T_{\text{ex}}$ inner iterations such that $T_{\text{tot}} \ge \frac{\log(\beta)}{\log \lambda_{\max}(\beta) \log(1-\pi)} \log(\epsilon)$.

Corollary 2 (Linear Convergence in the Primal)

Let $\hat{w}^t = -\frac{1}{\lambda}\Psi\hat{\mu}^t$. If we use SDCA on μ, then

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

Hence, if $\mathbb{E}[\hat{\Delta}_t + \Gamma_t]$ converges to 0 linearly, then so does $\mathbb{E}[P(\hat{w}^t) - P(w^*)]$.

8. Experiments

- *Baselines* using clique-wise updates:
  - SoftBCFW[2]/SoftSDCA: For a special case $\max_\mu D_\rho(\mu, \xi = 0)$.
  - GDMM[5]: Active-set ADMM-like algorithm.