Consistent Structured Prediction with Max-Min Markov Networks (M^{4}Ns)

Alex Nowak-Vila , Francis Bach and Alessandro Rudi

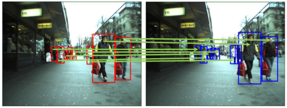
ICML 2020

Structured Output Prediction Setting

- Estimate $f : \mathcal{X} \to \mathcal{Y}$ that predicts *structured output* $y \in \mathcal{Y}$ from input $x \in \mathcal{X}$.

Handwritten Recognition

Matching



Structured Output Prediction Setting

1. **Prediction mistakes are not equally costly** $\rightarrow$ error measured with a loss $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and the goal is to solve

$$\underset{f:\mathcal{X}\rightarrow\mathcal{Y}}{\text{minimize}}\ \mathbb{E}\, L\big(f(x), y\big)$$

2. **The number of possible outputs is exponentially large** $\rightarrow$ encode structure with an embedding $\varphi : \mathcal{Y} \rightarrow \mathbb{R}^k$ with $k \ll |\mathcal{Y}|$.

$$f(x) = \underset{y\in\mathcal{Y}}{\arg\max}\ \varphi(y)^\top g(x).$$

Structured Output Prediction Setting

1. **Prediction mistakes are not equally costly** $\to$ error measured with a loss $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and the goal is to solve

$$\underset{f:\mathcal{X}\to\mathcal{Y}}{\text{minimize}} \ \mathbb{E}\, L\big(f(x), y\big)$$

2. **The number of possible outputs is exponentially large** $\to$ encode structure with an embedding $\varphi : \mathcal{Y} \to \mathbb{R}^k$ with $k \ll |\mathcal{Y}|$.

$$f(x) = \underset{y\in\mathcal{Y}}{\arg\max}\ \varphi(y)^\top g(x).$$

Problem to solve

$$\underset{g:\mathcal{X}\to\mathbb{R}^k}{\text{minimize}} \ \mathbb{E}\, L\Big(\underset{y\in\mathcal{Y}}{\arg\max}\ \varphi(y)^\top g(x), y\Big)$$

Structured Output Prediction Setting

1. **Prediction mistakes are not equally costly** $\rightarrow$ error measured with a loss $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and the goal is to solve

$$\underset{f:\mathcal{X}\longrightarrow\mathcal{Y}}{\text{minimize}} \ \mathbb{E}\, L\big(f(x), y\big)$$

2. **The number of possible outputs is exponentially large** $\rightarrow$ encode structure with an embedding $\varphi : \mathcal{Y} \rightarrow \mathbb{R}^k$ with $k \ll |\mathcal{Y}|$.

$$f(x) = \underset{y\in\mathcal{Y}}{\arg\max} \ \varphi(y)^\top g(x).$$

Problem to solve

$$\underset{g:\mathcal{X}\longrightarrow\mathbb{R}^k}{\text{minimize}} \ \mathbb{E}\, L\Big(\underset{y\in\mathcal{Y}}{\arg\max} \ \varphi(y)^\top g(x), y\Big)$$

It is non-convex! ✗

Max-Margin Markov Nets (M^{3}Ns) (a.k.a. SSVMs)

- **Convex upper bound** $\to$ Construct convex $S : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ s.t.

$$\mathbb{E}\, L\Big(\operatorname*{arg\,max}_{y \in \mathcal{Y}} \varphi(y)^\top g(x), y \Big) \le \mathbb{E}\, S(g(x), y)$$

Max-Margin Markov Nets (M^{3}Ns) (a.k.a. SSVMs)

- **Convex upper bound** $\rightarrow$ Construct convex $S : \mathbb{R}^k \times \mathcal{Y} \rightarrow \mathbb{R}$ s.t.

$$\mathbb{E}\, L\big(\arg\max_{y\in\mathcal{Y}}\; \varphi(y)^\top g(x), y\big) \leq \mathbb{E}\, S(g(x), y)$$

Max-Margin Markov Networks (M^{3}N)
([Taskar et al., 2004, Tsochantaridis et al., 2005])

$$S_{\mathrm{M^3N}}(v, y) = \max_{y'\in\mathcal{Y}}\; L(y, y') + v^\top \varphi(y') - v^\top \varphi(y),$$

Max-Margin Markov Nets (M^3Ns) (a.k.a. SSVMs)

- **Convex upper bound** $\to$ Construct convex $S : \mathbb{R}^k \times \mathcal{Y} \to \mathbb{R}$ s.t.

$$\mathbb{E}\, L\big(\arg\max_{y \in \mathcal{Y}} \varphi(y)^\top g(x), y\big) \leq \mathbb{E}\, S(g(x), y)$$

Max-Margin Markov Networks (M^3N)
([Taskar et al., 2004, Tsochantaridis et al., 2005])

$$S_{M^3N}(v, y) = \max_{y' \in \mathcal{Y}} L(y, y') + v^\top \varphi(y') - v^\top \varphi(y),$$

Inconsistent! ✗ ([Liu, 2007])

$$\lim_{n \to +\infty} \mathbb{E}\, S_{M^3N}(g_n(x), y) \longrightarrow \min_g \mathbb{E}\, S_{M^3N}(g(x), y)$$

$$\not\Longrightarrow$$

$$\lim_{n \to +\infty} \mathbb{E}\, L(\arg\max_{y' \in \mathcal{Y}} \varphi(y')^\top g_n(x), y) \longrightarrow \min_g \mathbb{E}\, L(\arg\max_{y' \in \mathcal{Y}} \varphi(y')^\top g(x), y)$$

Max-Min Margin Markov Networks (M^{4}Ns)

- M^{3}Ns can be re-written as:

$$S_{\mathrm{M^3N}}(v, y) = \max_{p \in \Delta_{\mathcal{Y}}} \; \mathbb{E}_{y' \sim p} L(y, y') + v^\top \varphi(y') - v^\top \varphi(y).$$

Max-Min Margin Markov Networks (M^{4}Ns)

- M^{3}Ns can be re-written as:

$$S_{\mathrm{M^3N}}(v, y) = \max_{p \in \Delta_{\mathcal{Y}}} \mathbb{E}_{y' \sim p} L(y, y') + v^\top \varphi(y') - v^\top \varphi(y).$$

Max-Min Margin Markov Networks (M^{4}N)

(based on [Fathony et al., 2018])

$$S_{\mathrm{M^4N}}(v, y) = \max_{p \in \Delta_{\mathcal{Y}}} \min_{z \in \mathcal{Y}} \mathbb{E}_{y' \sim p} L(z, y') + v^\top \varphi(y') - v^\top \varphi(y)$$

- **Not** an upper bound of L!
- For binary classification is **not** the SVM!

Contributions

Consistency ✓, Generalization Bound ✓

Contributions

Consistency ✓, Generalization Bound ✓

Algorithm for regularized ERM

- Based on BCFW [Lacoste-Julien et al., 2013] & Saddle-Point Mirror-Prox [Nemirovski, 2004].
- Requires **projection-oracle** instead of **max-oracle** of M^3Ns.

Contributions

Consistency ✓, Generalization Bound ✓

Algorithm for regularized ERM

- Based on BCFW [Lacoste-Julien et al., 2013] & Saddle-Point Mirror-Prox [Nemirovski, 2004].
- Requires **projection-oracle** instead of **max-oracle** of M^3Ns.

Statistical & Computational Guarantees of the Algorithm ✓

- **Setting:** regularized ERM in a RKHS $\mathcal{G}$.
- In the *worst case*, after $\boldsymbol{T = \mathcal{O}(n\sqrt{n})}$ **projections**, the output of the algorithm $\widehat{g}_{n,\boldsymbol{T}}$ satisfies

$$\mathbb{E}\, L\big(\underset{y' \in \mathcal{Y}}{\arg\max}\ \varphi(y')^\top \widehat{g}_{n,\boldsymbol{T}}(x), y\big) - \min_f \mathbb{E}\, L(f(x), y) \sim \|\varphi(f^\star)\|_{\mathcal{G}}\, n^{-1/2}.$$

Contributions

Consistency ✓, Generalization Bound ✓

Algorithm for regularized ERM

- Based on BCFW [Lacoste-Julien et al., 2013] & Saddle-Point Mirror-Prox [Nemirovski, 2004].
- Requires **projection-oracle** instead of **max-oracle** of M^{3}Ns.

Statistical & Computational Guarantees of the Algorithm ✓

- **Setting:** regularized ERM in a RKHS $\mathcal{G}$.
- In the *worst case*, after $\boldsymbol{T = \mathcal{O}(n\sqrt{n})}$ **projections**, the output of the algorithm $\widehat{g}_{n,\boldsymbol{T}}$ satisfies

$$\mathbb{E}\, L\Big(\underset{y' \in \mathcal{Y}}{\arg\max}\ \varphi(y')^\top \widehat{g}_{n,\boldsymbol{T}}(x), y\Big) - \min_f \mathbb{E}\, L(f(x), y) \sim \|\varphi(f^\star)\|_{\mathcal{G}}\, n^{-1/2}.$$

Experiments on sequences, matching and others. ✓

Plug-in (smooth) vs. Direct (non-smooth) Classifiers

$$g^\star = \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathbb{R}^k} \mathbb{E}\, S(g(x), y) \in \operatorname*{arg\,min}_{g:\mathcal{X}\to\mathbb{R}^k} \mathbb{E}\, L\big(\operatorname*{arg\,max}_{y\in\mathcal{Y}} \varphi(y)^\top g(x), y\big)$$

Plug-in (smooth) vs. Direct (non-smooth) Classifiers

$$g^\star = \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, S(g(x), y) \in \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, L\big(\underset{y\in\mathcal{Y}}{\arg\max}\, \varphi(y)^\top g(x), y\big)$$

1. Plug-in Classifiers (probabilistic)

- S is **smooth**.
- The moments $\mathbb{E}_{\boldsymbol{y'\sim\rho(\cdot|x)}}\, \boldsymbol{\varphi(y)}$ can be computed from $g^\star(x)$.

Plug-in (smooth) vs. Direct (non-smooth) Classifiers

$$g^\star = \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, S(g(x), y) \in \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, L\big(\underset{y\in\mathcal{Y}}{\arg\max}\, \varphi(y)^\top g(x), y\big)$$

1. Plug-in Classifiers (probabilistic)

- S is **smooth**.
- The moments $\mathbb{E}_{\boldsymbol{y'\sim\rho(\cdot|x)}}\, \boldsymbol{\varphi(y)}$ can be computed from $g^\star(x)$.

Binary classification ✓

$(\arg\max_{y'\in\mathcal{Y}} \varphi(y')^\top = \mathrm{sign},\ \mathbb{E}_{y'\sim\rho(\cdot|x)}\, \varphi(y) = \rho(1|x))$

- Examples: *Logistic* $\log(1+e^{-yv})$, *squared-hinge* $[1-yv]_+^2$.
- If $g^\star \in \mathcal{G}$ (RKHS). The regularized ERM $\widehat{g}_n$ satisfies

$$\mathbb{E}\, 1\big(\mathrm{sign}\circ\widehat{g}_n(x) \neq y\big) - \min_f \mathbb{E}\, 1(f(x)\neq y) \sim \|g^\star\|_{\mathcal{G}}\, n^{-1/4}.$$

Plug-in (smooth) vs. Direct (non-smooth) Classifiers

$$g^\star = \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, S(g(x), y) \in \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, L\Big(\underset{y\in\mathcal{Y}}{\arg\max}\, \varphi(y)^\top g(x), y\Big)$$

1. Plug-in Classifiers (probabilistic)

- S is **smooth**.
- The moments $\mathbb{E}_{\boldsymbol{y'\sim\rho(\cdot|x)}}\, \boldsymbol{\varphi(y)}$ can be computed from $g^\star(x)$.

Structured Prediction ✓

- Examples: *quadratic*, *conditional random fields* (CRF).

$$\tfrac{1}{2}\|v - \varphi(y)\|_2^2, \qquad \log\big(\textstyle\sum_{y'} \exp \varphi(y')^\top v\big) - v^\top \varphi(y).$$

- If $g^\star \in \mathcal{G}$ (RKHS). The regularized ERM $\widehat{g}_n$ satisfies

$$\mathbb{E}\, L\Big(\underset{y'\in\mathcal{Y}}{\arg\max}\ \varphi(y')^\top \widehat{g}_n(x), y\Big) - \min_f \mathbb{E}\, L(f(x), y) \sim \|g^\star\|_{\mathcal{G}}\, n^{-1/4}.$$

Plug-in (smooth) vs. Direct (non-smooth) Classifiers

$$g^\star = \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, S(g(x), y) \in \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, L\Big(\underset{y\in\mathcal{Y}}{\arg\max}\, \varphi(y)^\top g(x), y\Big)$$

2. Direct Classifiers

- S is **non-smooth**.
- $g^\star$ is **piece-wise constant.**

Plug-in (smooth) vs. Direct (non-smooth) Classifiers

$$g^\star = \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, S(g(x), y) \in \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, L\big(\underset{y\in\mathcal{Y}}{\arg\max}\, \varphi(y)^\top g(x), y\big)$$

2. Direct Classifiers

- S is **non-smooth**.
- $g^\star$ is **piece-wise constant.**

Binary classification ✓

$(\arg\max_{y'\in\mathcal{Y}} \varphi(y')^\top = \text{sign})$

- Examples: *binary SVM* $[1 - yv]_+$.
- If $f^\star \in \mathcal{G}$ (RKHS). The regularized ERM $\widehat{g}_n$ satisfies

$$\mathbb{E}\, 1\big(\text{sign} \circ \widehat{g}_n(x) \neq y\big) - \min_f \mathbb{E}\, 1(f(x) \neq y) \sim \|f^\star\|_{\mathcal{G}}\, n^{-1/2}.$$

Plug-in (smooth) vs. Direct (non-smooth) Classifiers

$$g^\star = \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, S(g(x), y) \in \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\, \mathbb{E}\, L\big(\underset{y\in\mathcal{Y}}{\arg\max}\, \varphi(y)^\top g(x), y\big)$$

2. Direct Classifiers

- S is **non-smooth**.
- $g^\star$ is **piece-wise constant.**

Structured Prediction

- Examples: *Max-Margin Markov Nets* ($\mathrm{M^3Ns}$) (a.k.a. SSVM).

$$S_{\mathrm{M^3N}}(v, y) = \max_{y'\in\mathcal{Y}}\; L(y, y') + v^\top\varphi(y') - v^\top\varphi(y)$$

- **Not consistent** ✗!

Plug-in (smooth) vs. Direct (non-smooth) Classifiers

$$g^\star = \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\,\mathbb{E}\,S(g(x),y) \in \underset{g:\mathcal{X}\to\mathbb{R}^k}{\arg\min}\,\mathbb{E}\,L\big(\underset{y\in\mathcal{Y}}{\arg\max}\,\varphi(y)^\top g(x), y\big)$$

2. Direct Classifiers

- S is **non-smooth**.
- $g^\star$ is **piece-wise constant.**

Structured Prediction

- Examples: *Max-Margin Markov Nets* (M[3]Ns) (a.k.a. SSVM).

$$S_{\mathrm{M^3N}}(v,y) = \max_{y'\in\mathcal{Y}}\ L(y,y') + v^\top\varphi(y') - v^\top\varphi(y)$$

- **Not consistent** ✗!

This paper: Complete the picture for Structured Prediction!

Related Work

1. **Smooth Surrogates.**
   - Quadratic [Ciliberto et al., 2016, Osokin et al., 2017].
   - Beyond quadratic [Nowak-Vila et al., 2019, Blondel, 2019].
2. **Non-smooth surrogates.**
   - Bounds on the ramp & margin loss.

     (consistent ✓, generalization bounds ✓, non-convex ✗)
     - PAC-Bayes bounds [Keshet and McAllester, 2011], [London et al., 2016].
     - Rademacher complexity bounds [Cortes et al., 2016]
   - Adversarial methods by [Fathony et al., 2016, Fathony et al., 2018, Duchi et al., 2018]

     (consistent ✓, generalization bounds ✗, convex ✓, no principled algorithm ✗)

From Max Margin to Max-Min Margin

$$S_{\mathrm{M^3N}}(v, y) = \underbrace{\max_{y' \in \mathcal{Y}} L(y, y') + v^\top \varphi(y')}_{\text{max oracle}} - v^\top \varphi(y)$$

$$S_{\mathrm{M^4N}}(v, y) = \underbrace{\max_{p \in \Delta_{\mathcal{Y}}} \min_{z \in \mathcal{Y}} \mathbb{E}_{y' \sim p} L(z, y') + v^\top \varphi(y')}_{\text{max-min oracle}} - v^\top \varphi(y)$$

From Max Margin to Max-Min Margin

$$S_{\mathrm{M^3N}}(v, y) = \underbrace{\max_{y' \in \mathcal{Y}} \; L(y, y') + v^\top \varphi(y')}_{\text{max oracle}} - v^\top \varphi(y)$$

$$S_{\mathrm{M^4N}}(v, y) = \underbrace{\max_{p \in \Delta_{\mathcal{Y}}} \min_{z \in \mathcal{Y}} \; \mathbb{E}_{y' \sim p} L(z, y') + v^\top \varphi(y')}_{\text{max-min oracle}} - v^\top \varphi(y)$$

Binary

M^3N	**M^4N**
$\max(1 - yv, 0)$	$\max(\lvert v \rvert, 1/2) - yv$

From Max Margin to Max-Min Margin

$$S_{\mathrm{M^3N}}(v, y) = \underbrace{\max_{y' \in \mathcal{Y}} \; L(y, y') + v^\top \varphi(y')}_{\text{max oracle}} - v^\top \varphi(y)$$

$$S_{\mathrm{M^4N}}(v, y) = \underbrace{\max_{p \in \Delta_{\mathcal{Y}}} \min_{z \in \mathcal{Y}} \; \mathbb{E}_{y' \sim p} L(z, y') + v^\top \varphi(y')}_{\text{max-min oracle}} - v^\top \varphi(y)$$

Binary

M³N	**M⁴N**
$\max(1 - yv, 0)$	$\max(\lvert v \rvert, 1/2) - yv$

Multi-Class

M³N	**M⁴N**
$\max_{j \neq y} 1 + v_j - v_y$	$1 + \max_{j \in [k]} \left\{ \frac{1}{j} \sum_{r=1}^{j} v_{(r)} - \frac{1}{j} \right\} - v_y$

with $v_{(1)} \geq \cdots \geq v_{(k)}$.

Completing the picture of Direct Classifiers

$$S_{\mathrm{M^3N}}(v, y) = \underbrace{\max_{y' \in \mathcal{Y}} \; L(y, y') + v^\top \varphi(y')}_{\text{max oracle}} - v^\top \varphi(y)$$

$$S_{\mathrm{M^4N}}(v, y) = \underbrace{\max_{p \in \Delta_{\mathcal{Y}}} \min_{z \in \mathcal{Y}} \; \mathbb{E}_{y' \sim p} \, L(z, y') + v^\top \varphi(y')}_{\text{max-min oracle}} - v^\top \varphi(y)$$

Statistical Properties of M⁴Ns ✓

- **Consistent.**
- If $\varphi(f^\star) \in \mathcal{G}$. The regularized ERM minimizer $\widehat{g}_n$ satisfies

$$\mathbb{E}\, L\Big(\operatorname*{arg\,max}_{y' \in \mathcal{Y}} \; \varphi(y')^\top \widehat{g}_n(x), y\Big) - \min_f \mathbb{E}\, L(f(x), y) \sim \frac{\|\varphi(f^\star)\|_{\mathcal{G}}}{n^{1/2}}.$$

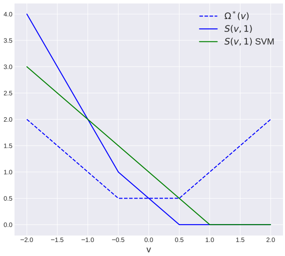
- The hidden constants in the bound are **not exponential**.

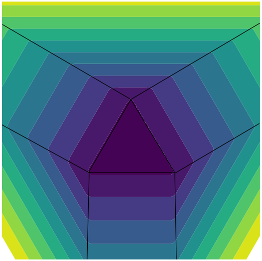
Images

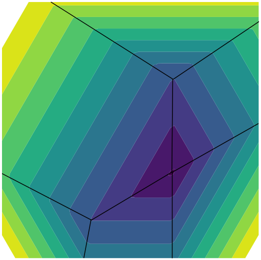
$$S_{\mathrm{M^4N}}(v, {\color{red}y}) = \underbrace{\max_{p \in \Delta_{\mathcal{Y}}} \min_{z \in \mathcal{Y}} \; \mathbb{E}_{y' \sim p} \, L(z, y') + v^\top \varphi(y')}_{{\color{blue}\Omega^*(v)}} - v^T \varphi({\color{red}y}).$$

- Function ${\color{blue}\Omega^*(v)}$ is a **non-smooth convex function**.
- Examples:

Binary	**Multi-class**	**Ordinal**
$L(y, y') = 1(y \neq y')$	$L(y, y') = 1(y \neq y')$	$L(y, y') = \lvert y - y' \rvert$







Algorithm for Regularized ERM

Problem: Computing the Regularized ERM

$$\widehat{g}_n = \min_g \frac{1}{n} \sum_{i=1}^{n} S_{\mathrm{M^4N}}(g(x_i), y_i) + \frac{\lambda}{2} \|g\|_{\mathcal{G}}^2$$

Algorithm for Regularized ERM

Problem: Computing the Regularized ERM

$\widehat{g}_n = \min_g \frac{1}{n} \sum_{i=1}^n S_{\mathrm{M^4N}}(g(x_i), y_i) + \frac{\lambda}{2} \|g\|_{\mathcal{G}}^2$

Computation of the Gradients

$$\arg\max_{p \in \Delta_{\mathcal{Y}}} \min_{z \in \mathcal{Y}} \; \mathbb{E}_{y' \sim p} L(z, y') + v^\top \varphi(y')$$

Algorithm for Regularized ERM

Problem: Computing the Regularized ERM

$\widehat{g}_n = \min_g \frac{1}{n} \sum_{i=1}^n S_{\mathrm{M^4N}}(g(x_i), y_i) + \frac{\lambda}{2} \|g\|_{\mathcal{G}}^2$

Computation of the Gradients

$$\arg\max_{p \in \Delta_{\mathcal{Y}}} \min_{z \in \mathcal{Y}} \ \mathbb{E}_{y' \sim p} L(z, y') + v^\top \varphi(y')$$

Approximate the gradients with (non-Euclidean) projections

Let $\mathcal{M} = \mathrm{hull}(\varphi(\mathcal{Y}))$ be the **marginal** polytope.

$$\arg\min_{\mu \in \mathcal{M}} \ v^\top \mu + H(\mu), \qquad H \text{ convex}$$

Guarantees of the Algorithm

Our oracle: (non-Euclidean) projections on $\mathcal{M} = \mathrm{hull}(\varphi(\mathcal{Y}))$

$$\arg\min_{\mu \in \mathcal{M}} \; v^\top \mu + H(\mu), \qquad H \text{ convex}$$

Statistical & Computational Guarantees of our Algorithm ✓

- Based on BCFW [Lacoste-Julien et al., 2013] & Saddle-Point Mirror-Prox [Nemirovski, 2004].
- In the **worst case**, after $\boldsymbol{T = \mathcal{O}(n\sqrt{n})}$ **projections**, the output of the algorithm $\widehat{g}_{n,\boldsymbol{T}}$ satisfies

$$\mathbb{E}\, L\Big(\arg\max_{y' \in \mathcal{Y}} \varphi(y')^\top \widehat{g}_{n,\boldsymbol{T}}(x), y\Big) - \min_f \mathbb{E}\, L(f(x), y) \sim \frac{\|\varphi(f^\star)\|_{\mathcal{G}}}{n^{1/2}}.$$

- In practice, $\boldsymbol{T = \mathcal{O}(n)}$ **projections** are enough using a **warm-start** strategy.

Max vs. Projection Oracle (Examples)

Sequence Prediction

- Sequence of length M and dictionary of size K.
- Hamming loss $L(y, y') = \frac{1}{M} \sum_{m=1}^{M} 1(y_m \neq y'_m)$.
- $\varphi(y)$ is a factor graph with unary and pairwise potentials.

max-oracle	**projection-oracle**
Max-product, $\mathcal{O}(MK^2)$	Sum-product, $\mathcal{O}(MK^2)$

Max vs. Projection Oracle (Examples)

Sequence Prediction

- Sequence of length M and dictionary of size K.
- Hamming loss $L(y, y') = \frac{1}{M}\sum_{m=1}^{M} 1(y_m \neq y'_m)$.
- $\varphi(y)$ is a factor graph with unary and pairwise potentials.

max-oracle	**projection-oracle**
Max-product, $\mathcal{O}(MK^2)$	Sum-product, $\mathcal{O}(MK^2)$

Matching

- M nodes to match.
- Hamming loss $L(\sigma, \sigma') = \frac{1}{M}\sum_{m=1}^{M} 1(\sigma(m) \neq \sigma'(m))$.
- $\varphi(\sigma)$ is the permutation matrix.

max-oracle	**projection-oracle**
Hungarian, $\mathcal{O}(M^3)$	Sinkhorn-Knopp, $\mathcal{O}(M^2/\varepsilon)$

Experiments

- Show effectiveness of M^4Ns compared to M^3Ns and CRFs on:

Multi-class Classification

Ordinal Regression

Sequence Prediction

Matching

Conclusion

- We introduced **Max-Min Markov Networks** (M^4Ns), a general method for structured prediction derived from first principles.
- We provide **consistency** guarantees and finite-sample **generalization bounds** on the regularized ERM analogous to the binary SVM.
- We provide an algorithm based on **non-Euclidean projections** that has both **computational and statistical guarantees**.
- We perform **experiments** on multiple structured prediction settings.

References

Blondel, M. (2019).
Structured prediction with projection oracles.
In *Advances in Neural Information Processing Systems*, pages 12145–12156.

Ciliberto, C., Rosasco, L., and Rudi, A. (2016).
A consistent regularization approach for structured prediction.
In *Advances in neural information processing systems*, pages 4412–4420.

Cortes, C., Kuznetsov, V., Mohri, M., and Yang, S. (2016).
Structured prediction theory based on factor graph complexity.
In *Advances in Neural Information Processing Systems*, pages 2514–2522.

Duchi, J., Khosravi, K., Ruan, F., et al. (2018).
Multiclass classification, information, divergence and surrogate risk.
The Annals of Statistics, 46(6B):3246–3275.

Fathony, R., Liu, A., Asif, K., and Ziebart, B. (2016).
Adversarial multiclass classification: A risk minimization perspective.
In *Advances in Neural Information Processing Systems*, pages