Book Chapter Supplement: Structured Learning from Cheap data

In this supplement we theoretically analyze the problem of learning with partially annotated outputs. We need a more refined notation as in the book chapter: we denote $\mathcal{Y}^{\circ}(y)$ and $\mathcal{Y}^{\circ}(y)$ for the space of all outputs that compatible and non-compatible, respectively, with y. Thus, for example, the term $\mathcal{Y}_n^{\circ}$ in the book chapter is denoted by $\mathcal{Y}^{\circ}(y_n)$ here. We also use a more little different notation for the loss l (see below).

We focus on the problem using the bridge loss, that is,

$$\min_{\boldsymbol{w}} \quad \lambda \left\|\boldsymbol{w}\right\|^2 + \frac{1}{N}\sum_{n=1}^{N}\left|\max_{y\in\mathcal{Y}^{\circ}(y_n)}\left(\langle \boldsymbol{w},\psi(x,y)\rangle + \Delta(y_n,y)\right) - \max_{y\in\mathcal{Y}^{*}(y_n)}\langle \boldsymbol{w},\psi(x,y)\rangle\right|_{+}.$$

First note that we can show by a standard Lagrangian argument (cf., e.g., Proposition 12 in [2]) that, for any $\lambda > 0$ there is an $\mu > 0$ such that the above problem can be equivalently rewritten as follows:

$$\min_{\boldsymbol{w}:\left\|\boldsymbol{w}\right\|\leq\mu} \quad \frac{1}{N}\sum_{n=1}^{N}\left|\max_{y\in\mathcal{Y}^{\circ}(y_n)}\left(\langle \boldsymbol{w},\psi(x,y)\rangle + \Delta(y_n,y)\right) - \max_{y\in\mathcal{Y}^{*}(y_n)}\langle \boldsymbol{w},\psi(x,y)\rangle\right|_{+}. \tag{1}$$

Now let us denote the base hypothesis class (the kernel class) by $\mathcal{F}^{\mathrm{ker}} := \left\{((x,y)\mapsto\langle \boldsymbol{w},\psi(x,y)\rangle) : \left\|\boldsymbol{w}\right\|\leq\mu\right\}$ and its induced structured-prediction class $\mathcal{F}^{\mathrm{struct}} := \left\{((x,y)\mapsto\rho_f(x,y)) : f\in\mathcal{F}^{\mathrm{ker}}\right\}$, where $\rho_f(x,y) := \max_{y'\in\mathcal{Y}^{\circ}(y)}\left(f(x,y')+\Delta(y,y')\right) - \max_{y'\in\mathcal{Y}^{*}(y)} f(x,y')$. Finally, denote the bridge-loss class by $\mathcal{G}^{\mathrm{bridge}} := l^{\mathrm{bridge}}\circ\mathcal{F}^{\mathrm{struct}}$, where $l^{\mathrm{bridge}}(t) := \left|t\right|_{+}$. Thus solving problem (1) is equivalent to performing *empirical risk minimization* over the class $\mathcal{G}_{\mathrm{bridge}}$, that is, $\min_{f\in\mathcal{F}^{\mathrm{struct}}}\frac{1}{N}\sum_{n=1}^{N} l^{\mathrm{bridge}}(f(x_n,y_n)) = \min_{g\in\mathcal{G}^{\mathrm{bridge}}}\frac{1}{N}\sum_{n=1}^{N} g(x_n,y_n)$. Hence, we may analyze structured prediction with partially annotation outputs within the proven framework of empirical risk minimization.

Background on Empirical Risk Minimization Let us briefly review the classic setup of empirical risk minimization [7]. We assume that $(x_1,y_1),\ldots,(x_N,y_N)$ is an i.i.d. sample drawn from a probability distribution P over $\mathcal{X}\times\mathcal{Y}$. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to some set $\mathcal{Y}$, and let $l:\mathcal{Y}\times\mathcal{Y}\to[0,b]$ be a loss function, for some $b>0$. The goal in statistical learning theory is to find a function $f\in\mathcal{F}$ that predicts well, i.e., that has a low risk $\mathbb{E}[l(f(x))]$. Denoting the loss class by $\mathcal{G} := l\circ\mathcal{F}$, this is equivalent finding a function g with small $\mathbb{E}[g]$. The best function in $\mathcal{G}$ we can hope to learn is $g^*\in\operatorname{argmin}_{g\in\mathcal{G}}\mathbb{E}[g]$.

Since g^* is unknown, we instead compute a minimizer $\widehat{g}_N\in\operatorname{argmin}_{g\in\mathcal{G}}\widehat{\mathbb{E}}[g]$, where $\widehat{\mathbb{E}}[g] := \frac{1}{N}\sum_{n=1}^{N} g(x_n)$. Let us compare the prediction accuracies of g^* and $\widehat{g}_N$. Standard learning theory gives [7] gives, with probability at least $1-\delta$ over the draw of the sample, $\mathbb{E}[\widehat{g}_N]-\mathbb{E}[g^*]\leq$

$$2\sup_{g\in\mathcal{G}}\left|\mathbb{E}[g]-\widehat{\mathbb{E}}[g]\right| \;\leq\; 2\mathbb{E}\sup_{g\in\mathcal{G}}\left|\mathbb{E}[g]-\widehat{\mathbb{E}}[g]\right| + b\sqrt{\frac{2\log(2/\delta)}{N}} \;\leq\; 4\mathfrak{R}_N(\mathcal{G}) + b\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{2}$$

The first inequality is a direct consequence of the minimality of $\widehat{f}$, the second one is by McDiamid's inequality [5], and the last inequality follows from symmetrization. Note that the result uses the notion of the *Rademacher complexity* $\mathfrak{R}_N(\mathcal{G})$, which is defined as follows.

Definition 1. *Let $\sigma_1, \ldots, \sigma_N$ be an i.i.d. family of Rademacher variables (random signs, i.e., each σ_i takes on the values -1 and 1, with equal probability of 1/2), independent of the sample $x_1, \ldots, x_N$. Then the* Rademacher complexity *of $\mathcal{G}$ is defined as $\mathfrak{R}_N(\mathcal{G}) := \mathbb{E} \sup_{g \in \mathcal{G}} \frac{1}{N} \sum_{n=1}^{N} \sigma_n g(x_n)$.*

Commonly $\mathfrak{R}_N(\mathcal{G})$ is of the order $O(1/\sqrt{N})$, when we employ appropriate regularization, so in that case the bound (2) converges at the order of $O(1/\sqrt{N})$. When bounding the Rademacher complexity for Lipschitz continuous loss classes (such as the hinge loss or the squared loss), the following lemma is often very helpful.

Lemma 2 (Talagrand's lemma; [4], Corollary 3.17)**.** *Let l be a loss function that is L-Lipschitz continuous and $l(0) = 0$. Let $\mathcal{F}$ be a hypothesis class of real-valued functions and denote its loss class by $\mathcal{G} := l \circ \mathcal{F}$. Then the following inequality holds: $\mathfrak{R}_N(\mathcal{G}) \leq 2L\mathfrak{R}_N(\mathcal{F})$.*

Generalization Guarantees for Structured Learning with Partially Annotated Outputs Let us denote the set of all possible partially annotated outputs by $\mathcal{Y}^p$. Then we have the following main theorem:

Theorem 3 (Generalization Bound for Structured Learning with Partially Annotated Outputs)**.** *Suppose there exist $b, B < \infty$ such that $\mathbb{P}\left(\sup_{g \in \mathcal{G}^{\mathrm{bridge}}} |g(x,y)| \leq b\right) = 1$ and $\mathbb{P}\left(\|\psi(x,y)\| \leq B\right) = 1$. Let $\Delta^{\max} := \sup_{y,y'} \Delta(y,y')$. Denote $g^* \in \operatorname{argmin}_{g \in \mathcal{G}^{\mathrm{bridge}}} \mathbb{E}[g]$ and $\widehat{g}_N \in \operatorname{argmin}_{g \in \mathcal{G}^{\mathrm{bridge}}} \widehat{\mathbb{E}}[g]$. Then, with probability at least $1 - \delta$, the generalization error of* structured prediction with partially annotated outputs *is bounded by:*

$$\mathbb{E}[\widehat{g}_N] - \mathbb{E}[g^*] \leq \frac{(\mu B + \Delta^{\max}) \left(8 \left|\mathcal{Y}^p\right| |\mathcal{Y}| + \sqrt{2 \log(2/\delta)}\right)}{\sqrt{N}}.$$

In particular, we have consistency, that is, $\mathbb{E}[\widehat{g}_N] - \mathbb{E}[g^*] \to 0$, when $N \to \infty$.

The proof, which is given below, uses similar ideas as in [3] for multi-class classification. In particular, the following result, taken from [6] (Lemma 8.1), is used.

Lemma 4. *Let $\mathcal{F}_1, \ldots, \mathcal{F}_l$ be hypothesis sets in $\mathbb{R}^{\mathcal{X}}$, and let $\mathcal{F} := \{\max(f_1, \ldots, f_l\} : f_i \in \mathcal{F}_i, i \in \{1, \ldots, l\}\}$. Then, $\mathfrak{R}_n(\mathcal{F}) \leq \sum_{j=1}^{l} \mathfrak{R}_n(\mathcal{F}_j)$.*

Proof. By assumption, we have almost surely,

$$\sup_{g \in \mathcal{G}^{\mathrm{bridge}}} |g(x,y)| \leq \sup_{f \in \mathcal{F}^{\mathrm{struct}}} |f(x,y)| \leq \sup_{f \in \mathcal{F}^{\ker}} |f(x,y)| + \Delta^{\max} \leq \mu B + \Delta^{\max}.$$

Thus, by (2), we have, with probability at least $1 - \delta$,

$$\mathbb{E}[\widehat{g}_N] - \mathbb{E}[g^*] \leq 4\mathfrak{R}_N(\mathcal{G}^{\mathrm{bridge}}) + (\mu B + \Delta^{\max}) \sqrt{\frac{2 \log(2/\delta)}{N}} \tag{3}$$

For the remainder of the proof it thus suffices to bound $\mathfrak{R}_N(\mathcal{G}^{\mathrm{bridge}})$. To this end, we proceed in three steps: 1. showing that $\mathfrak{R}_N(\mathcal{G}^{\mathrm{bridge}}) \leq 2\mathfrak{R}_N(\mathcal{F}^{\mathrm{struct}})$, 2. showing that $\mathfrak{R}_N(\mathcal{F}^{\mathrm{struct}})$ can be bounded by the term $\sum_{y^* \in \mathcal{Y}^p} \mathbb{E}\left[\sup_{f \in \mathcal{F}^{\ker}} \frac{1}{N} \sum_{n=1}^{N} \sigma_n \rho_f(x_i, y^*)\right]$, and 3. bounding the latter term.

STEP 1 The first step is also the simplest of the three: it is an obvious consequence of Talagrand's lemma (Lemma 2) as the bridge loss $l^{\mathrm{bridge}} : t \mapsto |t|_+$ is evidently 1-Lipschitz with $l^{\mathrm{bridge}}(0) := 0$.

STEP 2 Next, we note that

$$\begin{aligned}
\mathfrak{R}_N(\mathcal{F}^{\text{struct}}) &\overset{\text{def.}}{=} \mathbb{E}\left[\sup_{f\in\mathcal{F}^{\ker}} \frac{1}{N}\sum_{n=1}^{N}\sigma_n\rho_f(x_n,y_n)\right] \\
&\overset{(*)}{\leq} \sum_{y^*\in\mathcal{Y}^p}\mathbb{E}\left[\sup_{f\in\mathcal{F}^{\ker}} \frac{1}{N}\sum_{n=1}^{N}\sigma_n\rho_f(x_n,y_n)\mathbf{1}_{y^*=y_n}\right] \\
&\leq \frac{1}{2}\sum_{y^*\in\mathcal{Y}^p}\mathbb{E}\left[\sup_{f\in\mathcal{F}^{\ker}} \frac{1}{N}\sum_{n=1}^{N}\sigma_n\rho_f(x_n,y^*)\underbrace{(2\cdot\mathbf{1}_{y^*=y_n}-1)}_{\subseteq\{-1,1\}}\right] \\
&\quad + \frac{1}{2}\sum_{y^*\in\mathcal{Y}^p}\mathbb{E}\left[\sup_{f\in\mathcal{F}^{\ker}} \frac{1}{N}\sum_{n=1}^{N}\sigma_n\rho_f(x_n,y^*)\right] \\
&\overset{(**)}{=} \sum_{y^*\in\mathcal{Y}^p}\mathbb{E}\left[\sup_{f\in\mathcal{F}^{\ker}} \frac{1}{N}\sum_{n=1}^{N}\sigma_n\rho_f(x_n,y^*)\right],
\end{aligned}$$

where $(*)$ is by the sub-additivity of the supremum, and for $(**)$ we exploit that $-\sigma_n$ has the same distribution as σ_n.

STEP 3 We start by rewriting ρ_f explicitly:

$$\begin{aligned}
&\sum_{y^*\in\mathcal{Y}^p}\mathbb{E}\left[\sup_{f\in\mathcal{F}^{\ker}} \frac{1}{N}\sum_{n=1}^{N}\sigma_n\rho_f(x_n,y^*)\right] \\
&= \sum_{y^*\in\mathcal{Y}^p}\mathbb{E}\left[\sup_{f\in\mathcal{F}^{\ker}} \frac{1}{N}\sum_{n=1}^{N}\sigma_n \max_{y\in\mathcal{Y}^\circ(y^*)}(f(x,y)+\Delta(y^*,y)) - \max_{y\in\mathcal{Y}^*(y^*)} f(x,y)\right] \\
&\leq \sum_{y^*\in\mathcal{Y}^p}\left(\mathfrak{R}_N\left(\max_{y\in\mathcal{Y}^\circ(y^*)}\left(\left\{x\mapsto f(x,y)+\Delta(y^*,y) : f\in\mathcal{F}^{\ker}\right\}\right)\right)\right. \\
&\quad \left. + \mathfrak{R}_N\left(\max_{y\in\mathcal{Y}^*(y^*)}\left(\left\{x\mapsto f(x,y) : f\in\mathcal{F}^{\ker}\right\}\right)\right)\right)
\end{aligned}$$

We now may apply Lemma 4 to remove the maxima in the above bound, that is, for each y^*,

$$\mathfrak{R}_N\left(\max_{y\in\mathcal{Y}^*(y^*)}\left(\left\{x\mapsto f(x,y) : f\in\mathcal{F}^{\ker}\right\}\right)\right) \leq \sum_{y\in\mathcal{Y}^*(y^*)}\mathfrak{R}_N\left(\left(\left\{x\mapsto f(x,y) : f\in\mathcal{F}^{\ker}\right\}\right)\right)$$

and, similar,

$$\begin{aligned}
&\mathfrak{R}_N\left(\max_{y\in\mathcal{Y}^\circ(y^*)}\left(\left\{x\mapsto f(x,y)+\Delta(y^*,y) : f\in\mathcal{F}^{\ker}\right\}\right)\right) \\
&\leq \sum_{y\in\mathcal{Y}^\circ(y^*)}\mathfrak{R}_N\left(\left\{x\mapsto f(x,y)+\Delta(y^*,y) : f\in\mathcal{F}^{\ker}\right\}\right) \\
&\leq \sum_{y\in\mathcal{Y}^\circ(y^*)}\left(\mathfrak{R}_N\left(\left\{x\mapsto f(x,y) : f\in\mathcal{F}^{\ker}\right\}\right) + \underbrace{\Delta(y^*,y)}_{\leq\Delta^{\max}}/\sqrt{N}\right) \\
&\leq \frac{|\mathcal{Y}^\circ(y^*)|\,\Delta^{\max}}{\sqrt{N}} + \sum_{y\in\mathcal{Y}^\circ(y^*)}\mathfrak{R}_N\left(\left\{x\mapsto f(x,y) : f\in\mathcal{F}^{\ker}\right\}\right),
\end{aligned}$$

where we have used the well-known fact (e.g., Theorem 12.5 in [1]) that for a any constant $c\in\mathbb{R}$, $\mathfrak{R}_N(\mathcal{F}+c)\leq\mathfrak{R}_N(\mathcal{F})+|c|/\sqrt{N}$. Furthermore, the Rademacher complexity of kernel classes has been characterized in [1] (Lemma 22), which yields $\mathfrak{R}_N\left(\left(\left\{x\mapsto f(x,y) : f\in\mathcal{F}^{\ker}\right\}\right)\right) \leq \frac{\mu B}{\sqrt{N}}$. Thus, because $|\mathcal{Y}^\circ(y^*)|+|\mathcal{Y}^*(y^*)|=|\mathcal{Y}|$,

$$\sum_{y^*\in\mathcal{Y}^p}\mathbb{E}\left[\sup_{f\in\mathcal{F}^{\ker}} \frac{1}{N}\sum_{n=1}^{N}\sigma_n\rho_f(x_n,y^*)\right] \leq |\mathcal{Y}^p|\,|\mathcal{Y}|\,\frac{\mu B+\Delta^{\max}}{\sqrt{N}}.$$

Putting things together, we thus finally obtain the following bound on the Rademacher complexity:

$$\mathfrak{R}_N(\mathcal{G}^{\text{bridge}}) \overset{\text{STEP 1}}{\leq} 2\mathfrak{R}_N(\mathcal{F}^{\text{struct}}) \overset{\text{STEP 2}}{\leq} 2\sum_{y^* \in \mathcal{Y}^p} \mathbb{E}\left[\sup_{f \in \mathcal{F}^{\text{ker}}} \frac{1}{N}\sum_{n=1}^{N} \sigma_n \rho_f(x_n, y^*)\right]$$
$$\overset{\text{STEP 3}}{\leq} \frac{2\,|\mathcal{Y}^p|\,|\mathcal{Y}|\,(\mu B + \Delta^{\max})}{\sqrt{N}}\,.$$

Combining the above result with (3), we obtain the claimed result. □

References

[1] P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. 3:463–482, Nov. 2002.

[2] M. Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, Mar 2011.

[3] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

[4] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, New York, 1991.

[5] C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[6] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[7] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.