Sidestepping Intractable Inference with Structured Ensemble Cascades Supplementary Materials

David Weiss **Benjamin Sapp** **Ben Taskar**
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
{djweiss,bensapp,taskar}@cis.upenn.edu

A Video Clips

Included in the supplemental archive is a sample of test results `testset_samples.mp4`:

- The cyan lines indicate the best of the top $K = 4$ hypotheses for the ensemble; the red lines indicate the best of top $K = 4$ hypotheses for the single frame SC model.
- The point clouds show the top 150 unfiltered locations for each limb according to the ensemble model.
- The video illustrates the significant *qualitative* increase in accuracy of the ensemble over single-frame methods, eliminating the "jitter" due to the lack of smoothing.

B Proof of Theorem 1

We prove the theorem by applying the following result from [1]:

Theorem 1 (Bartlett and Mendelson, 2002)**.** *Consider a loss function $\mathcal{L}$ and a dominating cost function ϕ such that $\mathcal{L}(y, x) \leq \phi(y, x)$. Let $F : \mathcal{X} \mapsto \mathcal{A}$ be a class of functions. Then for any integer n and any $0 < \delta < 1$, with probability $1 - \delta$ over samples of length n, every f in F satisfies*

$$\mathbb{E}\mathcal{L}(Y, f(X)) \leq \hat{\mathbb{E}}_n\phi(Y, f(X)) + R_n(\tilde{\phi} \circ F) + \sqrt{\frac{8\ln(2/\delta)}{n}}, \tag{1}$$

where $\tilde{\phi} \circ F$ is a centered composition of ϕ with $f \in F$, $\tilde{\phi} \circ f = \phi(y, f(X)) - \phi(y, 0)$.

Furthermore, there are absolute constants c and C such that for every class F and every integer n,

$$cR_n(F) \leq G_n(F) \leq C \ln n R_n(F). \tag{2}$$

Let $\mathcal{A} = \mathbb{R}^m$ and $F : \mathcal{X} \to \mathcal{A}$ be a class of functions that is the direct sum of real-valued classes $F_1, \dots, F_m$. Then, for every integer n and every sample $(X_1, Y_1), \dots, (X_n, Y_n)$,

$$\hat{G}_n(\tilde{\phi} \circ F) \leq 2L \sum_{i=1}^{m} \hat{G}_n(F_i). \tag{3}$$

Let $F = \{x \mapsto \mathbf{w}^\top \mathbf{f}(x, \cdot) \mid ||\mathbf{w}||_2 \leq B, ||\mathbf{f}(x, \cdot)||_2 \leq 1\}$. Then,

$$\hat{G}_n(F) \leq \frac{2B}{\sqrt{n}}. \tag{4}$$

Proving the theorem reduces to analyzing the Lipschitz constant of the dominating cost function,

$$\phi(y, \theta_x) = r_\gamma \left(\frac{1}{P} \sum_p \theta_p(x, y) - t_p(x, \alpha) \right).$$

Let $\phi_p(y, \theta_p) = \theta_p(x, y) - t_p(x, \alpha)$. If we let θ_x^p be a m-dimensional vector whose elements correspond to the scores of every possible clique assignment given θ_p, then we can rewrite ϕ_p as the following:

$$\phi_p(y, \theta_x^p) = \langle y, \theta_x^p \rangle - t(x, \alpha),$$

where we consider y to be a binary m-dimensional vector that selects the active clique assignments in the output y. We now make use of the following lemma from [2]:

Lemma 1 (Weiss & Taskar, 2010)**.** *Let θ_x be a vector of clique assignment scores. Let $g(\theta_x) = \langle y, \theta_x \rangle - t(x, \alpha)$. Then $g(u) - g(v) \leq \sqrt{2\ell}||u - v||_2$.*

Lemma 2. *$\phi(y, \cdot)$ is Lipschitz with constant $\sqrt{2\ell}$.*

Proof. By Lemma 1, we see that ϕ_p is Lipschitz with constant $\sqrt{2\ell}$. Since ϕ is simply the average of P such functions, the Lipschitz of ϕ must be $\sqrt{2\ell}$ as well. □

Finally, to prove the theorem, we note that the loss function $\mathcal{L}(\theta, \langle X, Y \rangle, \alpha)$ can be represented as a function in $\mathbb{R}^{mP}$ space by contatenating the clique scoring vectors of each of the P individual sub-models. Substituing mP for the dimensionality of the loss function (m in Theorem 1) yields the desired result.

References

[1] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

[2] D. Weiss and B. Taskar. Structured prediction cascades. In *Proc. AISTATS*, 2010.