Supplementary material: Following the Perturbed Leader for Online Structured Learning

7. Appendix

7.1. Additional proofs

7.1.1. Maxima of normal random variables

Lemma 9. *Suppose the conditions of Theorem 1 hold, then*

$$\mathbb{E}\left[\max_{x\in\mathcal{X}}\langle x,\gamma\rangle\right] \le \sqrt{2k\log|\mathcal{X}|}$$

Proof. First, we upper bound the expectation by

$$\mathbb{E}_\gamma\left[\max_{x\in\mathcal{X}}\langle x,\gamma\rangle\right] \le \inf_{s>0}\frac{1}{s}\log\left(\sum_{x\in\mathcal{X}}\mathbb{E}[\exp(s\langle x,\gamma\rangle)]\right)$$

Notice that $\langle x,\gamma\rangle$ is a normal random variable with mean 0 and variance $\|x\|^2 \le k$. As such,

$$\mathbb{E}[\exp(s\langle x,\gamma\rangle)] = \exp\left(\frac{s^2\|x\|^2}{2}\right) \le \exp\left(\frac{ks^2}{2}\right)$$

Then,

$$\begin{aligned}\mathbb{E}_\gamma\left[\max_{x\in\mathcal{X}}\langle x,\gamma\rangle\right] &\le \inf_{s>0}\frac{1}{s}\log\left(|\mathcal{X}|\exp\left(\frac{ks^2}{2}\right)\right)\\ &= \inf_{s>0}\left\{\frac{\log|\mathcal{X}|}{s}+\frac{ks}{2}\right\}\\ &= \sqrt{2k\log|\mathcal{X}|}\end{aligned}$$

□

7.1.2. Bounding the Hessian

Lemma 10. *Suppose that the conditions of Theorem 1 hold. Let H denote the Hessian of Φ_η at an arbitrary θ. Fix some $j\in[d]$. Then,*

$$\sum_{i=1}^d|H_{i,j}| \le \frac{k}{\eta}\sum_{x\in\mathcal{X}}|\mathbb{E}[\gamma_j\mathbb{1}[\hat{x}=x]]|$$

Proof. Recall the definition of the Hessian:

$$H_{i,j} = \frac{1}{\eta}\mathbb{E}\left[\hat{x}(\tilde{\theta}_t+\eta\gamma)_i\gamma_j\right]$$

Let us abbreviate $\hat{x}(\theta+\eta\gamma)$ as $\hat{x}$. Then,

$$\begin{aligned}\eta\sum_{i=1}^d|H_{i,j}| &= \eta\sum_{i:H_{i,j}>0}H_{i,j} - \eta\sum_{i:H_{i,j}\le 0}H_{i,j}\\ &= \mathbb{E}\left[\left(\sum_{i:H_{i,j}>0}\hat{x}_i - \sum_{i:H_{i,j}\le 0}\hat{x}_i\right)\gamma_j\right]\\ &= \sum_{x\in\mathcal{X}}\mathbb{E}\left[\left(\sum_{i:H_{i,j}>0}\hat{x}_i - \sum_{i:H_{i,j}\le 0}\hat{x}_i\right)\gamma_j\mathbb{1}_{[\hat{x}=x]}\right]\\ &= \sum_{x\in\mathcal{X}}\left(\sum_{i:H_{i,j}>0}x_i - \sum_{i:H_{i,j}\le 0}x_i\right)\mathbb{E}\left[\gamma_j\mathbb{1}_{[\hat{x}=x]}\right]\\ &\le k\sum_{x\in\mathcal{X}}|\mathbb{E}[\gamma_j\mathbb{1}[\hat{x}=x]]|\end{aligned}$$

as $\left|\sum_{i:H_{i,j}>0}x_i - \sum_{i:H_{i,j}\le 0}x_i\right| \le k$ by assumption.

□

7.1.3. Bounding the Hessian for the k-sets problem

Proof of lemma 3. Let $H = \nabla^2\Phi_\eta(\tilde{\theta})$. We have that,

$$H_{i,j} = \frac{1}{\eta}\mathbb{E}\left[\hat{x}(\tilde{\theta}+\eta\gamma)_i\gamma_j\right]$$

with $\hat{x}(z)\in\arg\min_{x\in\mathcal{X}}\langle x,z\rangle$ (Abernethy et al., 2014, Lemma 7). We shall abbreviate $\hat{x}$ for $\hat{x}(\tilde{\theta}+\eta\gamma)$ in the remainder of the proof.

First, notice that

$$\sum_{i,j}H_{i,j} = \frac{1}{\eta}\sum_{i,j}\mathbb{E}[\gamma_j\hat{x}_i] = \frac{k}{\eta}\sum_{j=1}^d\mathbb{E}[\gamma_j] = 0$$

Secondly, we argue about the sign of $\mathbb{E}[\gamma_j\hat{x}_i]$. We claim that it is negative if $i=j$ and positive otherwise. To see that, notice that γ_j is a symmetric random variable, so that for each $\alpha>0$ the density of γ_j at α and at $-\alpha$ is the same. If $i\ne j$, the event $\hat{x}_i=1$ is more probable if $\gamma_j=\alpha$ than when $\gamma_j=-\alpha$. If $i=j$ then the opposite is true.

We have,

$$\begin{aligned}\sum_{i,j} H_{i,j} &= \sum_{i,j:H_{i,j}\geq 0} H_{i,j} - \sum_{i,j:H_{i,j}<0} H_{i,j} \\ &= -2 \sum_{i,j:H_{i,j}<0} H_{i,j} \\ &= -2\operatorname{Tr}(H)\end{aligned}$$

The rest of the proof follows that of lemma 2.

□

7.1.4. Technical lemma

Lemma 11. *We have,*

$$\begin{aligned}&\max\left\{\min\left\{\frac{Td}{16}, \frac{d\eta\sqrt{2}}{32}\right\}, \frac{Td}{16}\operatorname{erf}\left(-\frac{\sqrt{d}}{4\eta}\right)\right\} \\ &\quad\geq \min\left\{0.02Td, 0.05d^{5/4}\sqrt{T}\right\}\end{aligned}$$

Proof. We get,

$$\begin{aligned}&\max\left\{\min\left\{\frac{Td}{16}, \frac{d\eta\sqrt{2}}{32}\right\}, \frac{Td}{16}\operatorname{erf}\left(\frac{\sqrt{d}}{4\eta}\right)\right\} \\ &\quad\geq \min\left\{\frac{Td}{16},\right. \\ &\qquad\qquad \left.\max\left\{\frac{d\eta\sqrt{2}}{32}, \frac{Td}{16}\operatorname{erf}\left(\frac{\sqrt{d}}{4\eta}\right)\right\}\right\}\end{aligned}$$

Notice that erf is nondecreasing and concave on $\mathbb{R}_+$. Then,

$$\begin{aligned}&\inf_{\eta>0}\max\left\{\frac{d\eta\sqrt{2}}{32}, \frac{Td}{16}\operatorname{erf}\left(\frac{\sqrt{d}}{4\eta}\right)\right\} \\ &\quad\geq \min\left\{\inf_{\eta<\sqrt{d}/4}\frac{Td}{16}\operatorname{erf}\left(\frac{\sqrt{d}}{4\eta}\right),\right. \\ &\qquad\quad \left.\inf_{\eta\geq\sqrt{d}/4}\max\left\{\frac{d\eta\sqrt{2}}{32}, \frac{Td}{16}\operatorname{erf}\left(\frac{\sqrt{d}}{4\eta}\right)\right\}\right\} \\ &\quad\geq \min\left\{\frac{Td}{16}\operatorname{erf}(1),\right. \\ &\qquad\quad \left.\inf_{\eta\geq\sqrt{d}/4}\max\left\{\frac{d\eta\sqrt{2}}{32}, \frac{Td}{16}\frac{\sqrt{d}}{4\eta}\operatorname{erf}(1)\right\}\right\} \\ &\quad\geq \min\left\{\frac{Td}{16}\operatorname{erf}(1), \sqrt{\frac{d\sqrt{2}}{32}\frac{Td}{16}\frac{\sqrt{d}}{4}\operatorname{erf}(1)}\right\} \\ &\quad\geq \min\left\{0.05Td, 0.02d^{5/4}\sqrt{T}\right\}\end{aligned}$$

as required.

□

7.2. Lipschitz property of certain distributions

7.2.1. Uniform over the cube

Remember that we had required the marginals to have a variance of 1. Therefore WLOG we will take the cube to be $C = [0, 1/\sqrt{3}]^d$. Then,

$$\begin{aligned}\operatorname{TV}(P,Q) &= \sup_A \left|\Pr_P[A] - \Pr_Q[A]\right| \\ &= \sup_A \left|\frac{1}{\operatorname{Vol}(C)}\int_{x\in A} \mathbb{1}_{[x\in C+\{\mu_P\}]} - \mathbb{1}_{[x\in C+\{\mu_Q\}]}\right| \\ &\leq \sup_A \frac{1}{\operatorname{Vol}(C)}\int_{x\in A} \left|\mathbb{1}_{[x\in C+\{\mu_P\}]} - \mathbb{1}_{[x\in C+\{\mu_Q\}]}\right| \\ &\leq \frac{1}{\operatorname{Vol}(C)}\int_{x\in\mathbb{R}^d} \left|\mathbb{1}_{[x\in C+\{\mu_P\}]} - \mathbb{1}_{[x\in C+\{\mu_Q\}]}\right| \\ &= \frac{\operatorname{Vol}((C+\{\mu_P\}) \triangle (C+\{\mu_Q\}))}{\operatorname{Vol}(C)} \\ &\leq \frac{2(1/\sqrt{3})^{d-1}\|\mu_P-\mu_Q\|_1}{(1/\sqrt{3})^d} = 2\sqrt{3}\|\mu_P-\mu_Q\|_1\end{aligned}$$

so that $L = 2\sqrt{3}$.

We now explain the above bound. Suppose that $C+\{\mu_P\}$ and $C+\{\mu_Q\}$ do not intersect. Then we must have $\|\mu_P - \mu_Q\|_\infty > 1/\sqrt{3}$.

$$\begin{aligned}\operatorname{Vol}((C+\{\mu_P\}) \triangle (C+\{\mu_Q\})) &= 2\left(\frac{1}{\sqrt{3}}\right)^d \\ &< 2\left(\frac{1}{\sqrt{3}}\right)^{d-1}\|\mu_P-\mu_Q\|_\infty \\ &\leq 2\left(\frac{1}{\sqrt{3}}\right)^{d-1}\|\mu_P-\mu_Q\|_1\end{aligned}$$

If $C+\{\mu_P\}$ and $C+\{\mu_Q\}$ do intersect, then $\|\mu_P - \mu_Q\|_\infty \leq 1/\sqrt{3}$ and we have

$$\begin{aligned}&\operatorname{Vol}((C+\{\mu_P\}) \cap (C+\{\mu_Q\})) \\ &\quad= \prod_{i=1}^d \left(\frac{1}{\sqrt{3}} - |\mu_{P,i} - \mu_{Q,i}|\right)\end{aligned}$$

so that

$$
\begin{aligned}
&\mathrm{Vol}((C+\{\mu_P\}) \triangle (C+\{\mu_Q\})) \\
&\quad= \mathrm{Vol}(C+\{\mu_P\}) + \mathrm{Vol}(C+\{\mu_Q\}) \\
&\qquad - 2\mathrm{Vol}((C+\{\mu_P\}) \cap (C+\{\mu_Q\})) \\
&\quad= 2\left(\left(\frac{1}{\sqrt{3}}\right)^d - \prod_{i=1}^d \left(\frac{1}{\sqrt{3}} - |\mu_{P,i}-\mu_{Q,i}|\right)\right) \\
&\quad= 2\left(\frac{1}{\sqrt{3}}\right)^d \left(1 - \prod_{i=1}^d \left(1-\sqrt{3}|\mu_{P,i}-\mu_{Q,i}|\right)\right) \\
&\quad\le 2\left(\frac{1}{\sqrt{3}}\right)^d \sqrt{3}\|\mu_P-\mu_Q\|_1 \\
&\quad= 2\left(\frac{1}{\sqrt{3}}\right)^{d-1} \|\mu_P-\mu_Q\|_1
\end{aligned}
$$

7.2.2. Laplace and negative exponential

We will show that for the Laplace distribution we have $L = \sqrt{2}$. For the exponential distribution the proof is similar except with $L = 1$. Once again, recall that we had required the marginals to have a variance of 1, and therefore the PDF of the Laplace distribution is $\exp(-\sqrt{2}|x-\mu|)/\sqrt{2}$. In this case,

We want to bound

$$
\begin{aligned}
\mathrm{TV}(P,Q) &= \sup_A \left|\Pr_P[A] - \Pr_Q[A]\right| \\
&= \sup_A \left|\int_A \frac{1}{\sqrt{2}}\exp(-\sqrt{2}\|x-\mu_P\|_1)\right. \\
&\qquad \left. -\frac{1}{\sqrt{2}}\exp(-\sqrt{2}\|x-\mu_Q\|_1)\right|
\end{aligned}
$$

We have,

$$
\begin{aligned}
&\int_A \frac{1}{\sqrt{2}}\exp(-\sqrt{2}\|x-\mu_P\|_1) \\
&\qquad - \frac{1}{\sqrt{2}}\exp(-\sqrt{2}\|x-\mu_Q\|_1) \\
&\quad= \int_A \frac{1}{\sqrt{2}}\exp(-\sqrt{2}\|x-\mu_P\|_1) \\
&\qquad \cdot \left(1-\exp(\sqrt{2}\|x-\mu_P\|_1 - \sqrt{2}\|x-\mu_Q\|_1)\right) \\
&\quad\le \int_A \frac{1}{\sqrt{2}}\exp(-\sqrt{2}\|x-\mu_P\|_1) \\
&\qquad \cdot \left(\sqrt{2}\|x-\mu_Q\|_1 - \sqrt{2}\|x-\mu_P\|_1\right) \\
&\quad\le \int_A \frac{1}{\sqrt{2}}\exp(-\sqrt{2}\|x-\mu_P\|_1)\left(\sqrt{2}\|\mu_P-\mu_Q\|_1\right) \\
&\quad\le \sqrt{2}\|\mu_P-\mu_Q\|_1
\end{aligned}
$$

Similarly, one can bound

$$
\begin{aligned}
&\int_A \frac{1}{\sqrt{2}}\exp(-\sqrt{2}\|x-\mu_Q\|_1) \\
&\qquad - \frac{1}{\sqrt{2}}\exp(-\sqrt{2}\|x-\mu_P\|_1) \le \sqrt{2}\|\mu_P-\mu_Q\|_1
\end{aligned}
$$

and thus

$$
\mathrm{TV}(P,Q) \le \sqrt{2}\|\mu_P-\mu_Q\|_1
$$

as claimed.

References

Abernethy, Jacob, Lee, Chansoo, Sinha, Abhinav, and Tewari, Ambuj. Online linear optimization via smoothing. *The Journal of Machine Learning Research*, 35: 807–823, 2014.