REMARKS AND OMITTED PROOFS*

Abstract. We provide remarks and omitted proofs from "On the Connection Between Compression Learning and Scenario Based Single-Stage and Cascading Optimization Problems", Margellos, Prandini & Lygeros, IEEE Trans. Autom. Control, 60(10), 2015.

1. Proof of Theorem 1. The proof of this theorem is based on the proof of Theorem 5 in [1], and Theorem 1 in [7]. Without loss of generality fix $I_d = \{1, \dots, d\} \in \mathcal{I}_d$ and let $(\delta_1, \dots, \delta_d)$ in $\bar{\Delta}^d = \{(\delta_1, \dots, \delta_d) \in \Delta^d : \ d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\}$, where $\epsilon \in (0, 1)$. We have that

$$\begin{aligned} &\mathbb{P}\Big\{\delta \in \Delta : \ H_{I_d} \text{ is consistent with } \big(\delta, \mathbb{1}_T(\delta)\big) \text{ and } d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\Big\} \\ &\qquad = \mathbb{P}\Big\{\delta \in \Delta : \ H_{I_d} \text{ is consistent with } \big(\delta, \mathbb{1}_T(\delta)\big)\Big\} = 1 - d_{\mathbb{P}}(T, H_{I_d}) \leq 1 - \epsilon, \end{aligned} \tag{1.1}$$

where the first step follows from the fact that $d_{\mathbb{P}}(T, H_{I_d})$ does not depend on δ but only on $(\delta_1, \dots, \delta_d) \in \bar{\Delta}^d$, and the second step follows from the definition of a consistent hypothesis (Definition 2). In fact, the inequality in (1.1) is strict as $d_{\mathbb{P}}(T, H_{I_d}) > \epsilon$. Since the samples are extracted independently we have that

$$\begin{aligned} &\mathbb{P}^{m-d}\Big\{(\delta_{d+1}, \dots, \delta_m) \in \Delta^{m-d} : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=d+1}^m \text{ and } d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\Big\} \\ &\quad = \mathbb{P}^{m-d}\Big\{(\delta_{d+1}, \dots, \delta_m) \in \Delta^{m-d} : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=d+1}^m\Big\} \\ &\quad = \prod_{j=d+1}^{m} \mathbb{P}\Big\{\delta_j \in \Delta : \ H_{I_d} \text{ is consistent with } \big(\delta_j, \mathbb{1}_T(\delta_j)\big)\Big\} \leq (1-\epsilon)^{m-d}, \end{aligned} \tag{1.2}$$

where the first equality is due to the fact that $d_{\mathbb{P}}(T, H_{I_d})$ does not depend on $\delta_{d+1}, \dots, \delta_m$, and since by the theorem's hypothesis, H_{I_d} is consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_d}$. However, (1.2) can be rewritten by the following conditional probability as $\bar{\Delta}^d$ is a cylinder set with base the cartesian product of the domains of the first d multi-sample elements (since it is independent of $\{\delta_i\}_{i=d+1}^m$), i.e.,

$$\begin{aligned} &\mathbb{P}^m\Big\{\big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big\} \ \Big| \ \big\{(\delta_1, \dots, \delta_d) \in \bar{\Delta}^d\big\}\Big\} \\ &\qquad\qquad \leq (1-\epsilon)^{m-d}. \end{aligned} \tag{1.3}$$

Deconditioning, we get

$$\begin{aligned} &\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\Big\} \\ &\quad \leq \int_{\Delta^d} (1-\epsilon)^{m-d} \mathbb{1}_{\big\{(\delta_1, \dots, \delta_d) \in \bar{\Delta}^d\big\}} \mathrm{d}\mathbb{P}^d\big(\{\delta_i\}_{i \in I_d}\big) \leq (1-\epsilon)^{m-d}. \end{aligned} \tag{1.4}$$

The left-hand side of (2) in the statement of Theorem 1 can be then expressed as follows

$$\begin{aligned} &\mathbb{P}^m\Big\{\bigcup_{I_d \in \mathcal{I}_d} \big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\big\}\Big\} \\ &\quad \leq \sum_{I_d \in \mathcal{I}_d} \mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\Big\} \\ &\quad \leq \binom{m}{d}(1-\epsilon)^{m-d}, \end{aligned} \tag{1.5}$$

concluding the proof (the first inequality is by the subadditivity of $\mathbb{P}^m$ and the second one due to (1.4)). If $\bar{\Delta}^d$ was empty, then the corresponding term in the summation would be zero, thus bounded by $(1-\epsilon)^{m-d}$.

*28.01.2022

2. Proof of Theorem 3. The proof of this theorem is based on the first part of the proof of Theorem 1 in [8]. Without loss of generality fix $I_d = \{1, \dots, d\} \in \mathcal{I}_d$. Let $V : \Delta^d \to [0,1]$ be a random variable such that $V(\{\delta_i\}_{i=1}^d) = d_{\mathbb{P}}(T, H_{I_d})$ for all $\{\delta_i\}_{i=1}^d \in \Delta^d$. For $v \in [0,1]$ denote its probability distribution function by $F(v) = \mathbb{P}^d\Big\{\{\delta_i\}_{i=1}^d \in \Delta^d : \ d_{\mathbb{P}}(T, H_{I_d}) \le v\Big\} = \mathbb{P}^d\Big\{V^{-1}\big([0,v]\big)\Big\} = \mathbb{P}_V^d\Big\{V \le v\Big\}$, where $\mathbb{P}_V^d$ denotes the image probability of $\mathbb{P}^d$ through V, defined over the Borel σ-algebra on $[0,1]$ (see Chapter 2 in [Campi, Selected Topics in Probability, Lecture Notes, 2008]). We used that, for $a < b$, $V^{-1}\big([a,b]\big) = \big\{\{\delta_i\}_{i=1}^d \in \Delta^d : \ V(\{\delta_i\}_{i=1}^d) \in [a,b]\big\}$.

By the first part of Assumption 3 and due to the fact that for a fixed $(\delta_1, \dots, \delta_d) \in \Delta^d$, $d_{\mathbb{P}}(T, H_{I_d}) = V(\{\delta_i\}_{i=1}^d)$, (1.3) becomes

$$\mathbb{P}^m\Big\{\big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big\} \ \Big| \ \big\{(\delta_1, \dots, \delta_d) \in \Delta^d\big\}\Big\} = \Big(1 - V(\{\delta_i\}_{i=1}^d)\Big)^{m-d}. \tag{2.1}$$

Deconditioning and using the equivalences between a probability and its image, (2.1) yields then

$$\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\Big\} = \int_{\Delta^d} \Big(1 - V(\{\delta_i\}_{i=1}^d)\Big)^{m-d} \, \mathrm{d}\mathbb{P}^d\big(\{\delta_i\}_{i \in I_d}\big) = \int_0^1 (1-v)^{m-d} \, \mathrm{d}F(v). \tag{2.2}$$

By Assumption 3 the sets $\Big\{\big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big\}\Big\}_{I_d \in \mathcal{I}_d}$ form a partition of Δ^m up to a set of measure zero[1]. Since there are $\binom{m}{d}$ sets in $\mathcal{I}_d$, we have that $\binom{m}{d}$ $\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\Big\} = 1$, which due to (2.2) leads to

$$\binom{m}{d} \int_0^1 (1-v)^{m-d} \, \mathrm{d}F(v) = 1, \text{ for all } m \ge d \implies F(v) = v^d, \tag{2.3}$$

where the $F(v)$ above is the unique solution satisfying the previous identities. This is due to [8], based on the fact that this is a moment problem for a distribution with finite support ($[0,1]$ here); see, e.g., Chapter 2, Section 12.9, Corollary 1 of [A. Shiryaev, Probability, Springer, 1996]. It can be verified using integration by parts. In place of (1.4) we thus have

$$\begin{aligned}
&\mathbb{P}^m\Big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m \text{ and } d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\Big\} \\
&= \int_{\Delta^d} \mathbb{P}^m\Big\{\big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big\} \\
&\qquad\qquad \Big| \ \big\{(\delta_1, \dots, \delta_d) \in \Delta^d : \ d_{\mathbb{P}}(T, H_{I_d}) > \epsilon\big\}\Big\} \ \mathbb{1}_{\left\{V^{-1}\left((\epsilon,1]\right)\right\}} \, \mathrm{d}\mathbb{P}^d\big(\{\delta_i\}_{i \in I_d}\big) \\
&= \int_{\Delta^d} \Big(1 - V(\{\delta_i\}_{i=1}^d)\Big)^{m-d} \mathbb{1}_{\left\{V^{-1}\left((\epsilon,1]\right)\right\}} \, \mathrm{d}\mathbb{P}^d\big(\{\delta_i\}_{i \in I_d}\big) \\
&= \int_\epsilon^1 (1-v)^{m-d} \mathrm{d}F(v) = \int_\epsilon^1 (1-v)^{m-d} dv^{d-1} \mathrm{d}v,
\end{aligned} \tag{2.4}$$

where we used (2.1)[2], since we still condition on some fixed $\delta_1, \dots, \delta_d$, and (2.3) as $\mathrm{d}F(v) = dv^{d-1} \, \mathrm{d}v$.

---

[1] To see this, for all $I_d \in \mathcal{I}_d$ let $S_{I_d} = \big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big\}$. We first show that $\mathbb{P}^m\big\{(\delta_1, \dots, \delta_m) \in \Delta^m : \ \Delta^m \backslash \cup_{I_d \in \mathcal{I}_d} S_{I_d}\big\} = 0$. Equivalently, $\cup_{I_d \in \mathcal{I}_d} S_{I_d} = \Delta^m$ up to a set of measure zero. Since $\cup_{I_d \in \mathcal{I}_d} S_{I_d} \subseteq \Delta^m$, we show that $\cup_{I_d \in \mathcal{I}_d} S_{I_d} \supseteq \Delta^m$, i.e. if $(\delta_1, \dots, \delta_m) \in \Delta^m$ then there exists $I_d \in \mathcal{I}_d$ such that $(\delta_1, \dots, \delta_m) \in S_{I_d}$. With $\mathbb{P}^m$-probability one, the last statement follows from Assumption 3 and the definition of S_{I_d}. We now show that $S_{I_d^1} \cap S_{I_d^2} = \emptyset$ for all $I_d^1, I_d^2 \in \mathcal{I}_d$ with $I_d^1 \ne I_d^2$. For the sake of contradiction assume that there exist $I_d^1, I_d^2 \in \mathcal{I}_d$ with $I_d^1 \ne I_d^2$ such that $S_{I_d^1} \cap S_{I_d^2} \ne \emptyset$. By the definition of $S_{I_d^1}, S_{I_d^2}$, this implies that there exists $(\delta_1, \dots, \delta_m) \in \Delta^m$ such that both $H_{I_d^1}$ and $H_{I_d^2}$ are consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m$. However, by Assumption 3 a unique hypothesis consistent with respect to the m-multisample should exist almost surely, establishing a contradiction up to a measure zero set.

[2] Alternatively, we could see this by setting $A = \{(\delta_1, \dots, \delta_m) \in \Delta^m : \ H_{I_d} \text{ is consistent with } \big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\}$, $B = (\epsilon, 1]$, and writing (2.4) as $\mathbb{P}^m\{A \cap \{V \in B\}\} = \int_B \mathbb{P}^m\{A | V = v\} \, \mathrm{d}F(v)$ (see [A. Shiryaev, Probability, Springer, 1996], Chapter 2, Section 7.5, eq. (17)). Noticing then that due to (2.1), $\mathbb{P}^m\{A | V = v\} = (1-v)^{m-d}$ the last statement yields the result in (2.4).

Notice now that the first inequality in (1.5) holds with equality since the union in the left-hand side of (1.5) is disjoint as by Assumption 3 the associated sets form a partition of Δ^m almost surely. We thus have

$$\begin{aligned}(2.5)\qquad &\mathbb{P}^m\Big\{\bigcup_{I_d\in\mathcal{I}_d}\big\{(\delta_1,\dots,\delta_m)\in\Delta^m:\ H_{I_d}\text{ is consistent with }\big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\text{ and }d_{\mathbb{P}}(T,H_{I_d})>\epsilon\big\}\Big\}\\ &\quad=\binom{m}{d}\int_\epsilon^1(1-v)^{m-d}dv^{d-1}\mathrm{d}v=\sum_{i=0}^{d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i},\end{aligned}$$

where the last equality follows by repeated integration by parts [8], and is the cumulative distribution of a binomial random variable. By the second part of Assumption 3, there exists a unique $m_d\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\in\mathcal{I}_d$ such that H_{m_d} is a consistent hypothesis. Eq. (2.5) is thus $\mathbb{P}^m\big\{(\delta_1,\dots,\delta_m)\in\Delta^m:\ d_{\mathbb{P}}(T,H_{m_d})>\epsilon\big\}$, concluding the proof.

3. Proof of Theorem 4. The proof of this theorem is based on the proof of Theorem 2.1 in [20]. Fix any $r\in\mathbb{N}$ and $I_r\in\mathcal{I}_r$, and let $m_d^r\Big(\big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i\in\{1,\dots,m\}\setminus I_r}\Big)$ be a set of d indices satisfying the first two parts of Assumption 4. Letting $V\big(\{\delta_i\}_{i\in\{1,\dots,m\}\setminus I_r}\big)=d_{\mathbb{P}}(T,H_{m_d^r})$, we have

$$(3.1)\qquad \mathbb{P}\Big\{\delta\in\Delta:\ H_{m_d^r}\text{ is not consistent with }\big(\delta,\mathbb{1}_T(\delta)\big)\Big\}=d_{\mathbb{P}}(T,H_{m_d^r})=V\big(\{\delta_i\}_{i\in\{1,\dots,m\}\setminus I_r}\big).$$

Since the samples are extracted independently we have that

$$\begin{aligned}(3.2)\qquad &\mathbb{P}^r\Big\{\{\delta_j\}_{j\in I_r}\in\Delta^r:\ H_{m_d^r}\text{ is not consistent with }\big\{\big(\delta_j,\mathbb{1}_T(\delta_j)\big)\big\}_{j\in I_r}\Big\}\\ &\quad=\prod_{j\in I_r}\mathbb{P}\Big\{\delta_j\in\Delta:\ H_{m_d^r}\text{ is not consistent with }\big(\delta_j,\mathbb{1}_T(\delta_j)\big)\Big\}=V\big(\{\delta_i\}_{i\in\{1,\dots,m\}\setminus I_r}\big)^r.\end{aligned}$$

Let $\bar F(v)=\mathbb{P}^{m-r}\big\{\{\delta_i\}_{i\in\{1,\dots,m\}\setminus I_r}\in\Delta^{m-r}:\ d_{\mathbb{P}}(T,H_{m_d^r})\le v\big\}$ denote the distribution function of V. Similarly to (2.4) in the proof of Theorem 3, we then have that

$$\begin{aligned}(3.3)\qquad &\mathbb{P}^m\Big\{(\delta_1,\dots,\delta_m)\in\Delta^m:\ H_{m_d^r}\text{ is not consistent with }\big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_r}\text{ and }d_{\mathbb{P}}(T,H_{m_d^r})>\epsilon\Big\}\\ &\quad=\int_\epsilon^1\mathbb{P}^m\Big\{\big\{(\delta_1,\dots,\delta_m)\in\Delta^m:\ H_{m_d^r}\text{ is not consistent with }\big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_r}\big\}\\ &\qquad\qquad\Big|\ V\big(\{\delta_i\}_{i\in\{1,\dots,m\}\setminus I_r}\big)=v\Big\}\,\mathrm{d}\bar F(v)\\ &\quad=\int_\epsilon^1 v^r\mathrm{d}\bar F(v).\end{aligned}$$

Construct the algorithm $\big\{A_{m-r}\big\}_{m-r\ge d}$, such that $A_{m-r}\big(\big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i\in\{1,\dots,m\}\setminus I_r}\big)=H_{m-r}=H_{m_d^r}$. By Theorem 1, since m_d^r satisfies the second part of Assumption 4, with $m-r$ in place of m and v in place of ϵ, we have that the constructed algorithm is PAC-T, hence for any $v\in(0,1)$,

$$\begin{aligned}(3.4)\qquad &\mathbb{P}^{m-r}\Big\{\{\delta_i\}_{i\in\{1,\dots,m\}\setminus I_r}\in\Delta^{m-r}:\ d_{\mathbb{P}}(T,H_{m_d^r})>v\Big\}\le\binom{m-r}{d}(1-v)^{m-r-d}\\ &\implies\bar F(v)\ge F(v):=1-\binom{m-r}{d}(1-v)^{m-r-d}.\end{aligned}$$

We then have that

$$\begin{aligned}(3.5)\qquad \int_\epsilon^1 v^r\mathrm{d}\bar F(v)\le\int_\epsilon^1 v^r\mathrm{d}F(v)&=\int_\epsilon^1(m-r-d)\binom{m-r}{d}v^r(1-v)^{m-r-d-1}\mathrm{d}v\\ &=\binom{m-r}{d}\frac{1}{\binom{m-d}{r}}\sum_{i=0}^r\binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i},\end{aligned}$$

where the first equality is due to the definition of $F(v)$, and the second one follows by repeated integration by parts. To see the inequality, notice that

$$\begin{aligned}(3.6)\qquad \int_\epsilon^1 v^r\mathrm{d}\bar F(v)&=1-\epsilon^r\bar F(\epsilon)-\int_\epsilon^1\bar F(v)rv^{r-1}\mathrm{d}v\\ &\le 1-\epsilon^r F(\epsilon)-\int_\epsilon^1 F(v)rv^{r-1}\mathrm{d}v=\int_\epsilon^1 v^r\mathrm{d}F(v),\end{aligned}$$

where the equality follows from Theorem 11, Chapter 2, Section 6.11 of [A. Shiryaev, Probability, Springer, 1996], since v^r is an increasing function of v, hence it is treated as a generalized distribution (see also [20]). The inequality is since $\bar{F}(v) \geq F(v)$ by (3.4). Intuitively, $\bar{F}(v) \geq F(v)$ is a dominance condition for cumulative distributions, implying that F is concentrated to higher values compared to $\bar{F}$ (if they admit a density, the density of F is also concentrated to higher values); hence, the associated expected value ($\int_\epsilon^1 v^r \mathrm{d}F(v)$ is related to the expected value of an increasing function of v) would be higher compared to the one corresponding to $\bar{F}$.

Consider the third part of Assumption 4. Denote then by $\bar{I}_r \in \mathcal{I}_r$ the associated set of indices, and let $\bar{m}_d^r\Big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\{1,\ldots,m\}\setminus\bar{I}_r}\Big)$ be a set of d indices such that $H_{\bar{m}_d^r} = G_d\big(\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\bar{m}_d^r}\big)$ is not consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i\in\bar{I}_r}$. Since this happens with $\mathbb{P}^m$-probability one, we have the first inequality (this is not equality since $\bar{I}_r$ is not necessarily unique) below, i.e.,

$$
\begin{aligned}
(3.7)\quad & \mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m)\in\Delta^m:\ d_{\mathbb{P}}(T, H_{\bar{m}_d^r}) > \epsilon\Big\} \\
& \leq \mathbb{P}^m\Big\{\bigcup_{I_r\in\mathcal{I}_r}\big\{(\delta_1,\ldots,\delta_m)\in\Delta^m:\ H_{m_d^r} \text{ is not consistent with } \big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_r} \text{ and } d_{\mathbb{P}}(T,H_{m_d^r})>\epsilon\big\}\Big\} \\
& \leq \binom{m}{r}\binom{m-r}{d}\frac{1}{\binom{m-d}{r}}\sum_{i=0}^{r}\binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i} = \binom{m}{d}\sum_{i=0}^{r}\binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i},
\end{aligned}
$$

where the second inequality is by the subadditivity of $\mathbb{P}^m$ since there are $\binom{m}{r}$ sets in the union, and by (3.3) and (3.5). Set $q(m,\epsilon) = \binom{m}{d}\sum_{i=0}^{r}\binom{m-d}{i}\epsilon^i(1-\epsilon)^{m-d-i}$. Similarly to the last part of the proof of Theorem 2, $\lim_{m\to\infty} q(m,\epsilon) = 0$. Construct then algorithm $\big\{A_m\big\}_{m\geq d+r}$, where $A_m : [\Delta\times\{0,1\}]^m \to \mathcal{D}$ takes as input a labeled m-multisample and returns a hypothesis $H_m = A_m\big(\big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i=1}^m\big)$ such that $H_m = H_{\bar{m}_d^r}$. By Definition 4, algorithm $\big\{A_m\big\}_{m\geq d+r}$ is PAC-T, thus concluding the proof.

4. Tightening Theorem 4. If we strengthen Assumption 4 by further requiring that the set $I_d \in \mathcal{I}_d^{m-r}$ that satisfies its conditions is unique, we could replace the right-hand side of (3.7) with $\binom{r+d-1}{r}\sum_{i=0}^{r+d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i}$. The proof of this fact is identical to the one of Theorem 4, with the difference that (3.4) is replaced by $\bar{F}(v) = F(v) := 1-\sum_{i=0}^{d-1}\binom{m-r}{i}v^i(1-v)^{m-r-i}$. Eq. (3.5) would then hold with equality, however, the final result would still hold with inequality due to the inequalities in (3.7).

Specializing this in a convex scenario optimization context, adopt an algorithm that removes the samples with indices in $\bar{I}_r$, and generates a hypothesis $H_{\bar{m}_d^r}$ such that $d_{\mathbb{P}}(T, H_{\bar{m}_d^r})$ is the probability of constraint violation. We further require that $H_{\bar{m}_d^r}$ is inconsistent with the removed samples (i.e., the associated minimizer violates the constraints for the removed samples), and for any given $(\delta_1,\ldots,\delta_m)$, $d_{\mathbb{P}}(T, H_{\bar{m}_d^r}) \leq d_{\mathbb{P}}(T, H_{m_d^r})$ for any set I_r in the family of sets of inconsistent hypotheses[3]. In words, we consider a conservative design, removing samples that lead to the lowest probability of constraint violation. We then have that,

$$
\begin{aligned}
(4.1)\quad & \binom{r+d-1}{r}\,\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m)\in\Delta^m:\ d_{\mathbb{P}}(T,H_{\bar{m}_d^r})>\epsilon\Big\} \\
& = \int_{\Delta^m}\mathbb{1}_{\left\{(\delta_1,\ldots,\delta_m):\ d_{\mathbb{P}}(T,H_{\bar{m}_d^r})>\epsilon\right\}}\sum_{I_r\in\mathcal{I}_r}\mathbb{1}_{\left\{(\delta_1,\ldots,\delta_m):\ H_{m_d^r}\text{ is not consistent with }\left\{\left(\delta_i,\mathbb{1}_T(\delta_i)\right)\right\}_{i\in I_r}\right\}}\,\mathrm{d}\mathbb{P}^m\big(\{\delta_i\}_{i=1}^m\big) \\
& \leq \int_{\Delta^m}\sum_{I_r\in\mathcal{I}_r}\mathbb{1}_{\left\{(\delta_1,\ldots,\delta_m):\ H_{m_d^r}\text{ is not consistent with }\left\{\left(\delta_i,\mathbb{1}_T(\delta_i)\right)\right\}_{i\in I_r}\right\}}\mathbb{1}_{\left\{(\delta_1,\ldots,\delta_m):\ d_{\mathbb{P}}(T,H_{m_d^r})>\epsilon\right\}}\,\mathrm{d}\mathbb{P}^m\big(\{\delta_i\}_{i=1}^m\big) \\
& = \sum_{I_r\in\mathcal{I}_r}\mathbb{P}^m\Big\{(\delta_1,\ldots,\delta_m)\in\Delta^m:\ H_{m_d^r}\text{ is not consistent with }\big\{\big(\delta_i,\mathbb{1}_T(\delta_i)\big)\big\}_{i\in I_r}\text{ and }d_{\mathbb{P}}(T,H_{m_d^r})>\epsilon\Big\} \\
& = \binom{r+d-1}{r}\sum_{i=0}^{r+d-1}\binom{m}{i}\epsilon^i(1-\epsilon)^{m-i},
\end{aligned}
$$

[3]For the class of fully-supported scenario programs, for each $(\delta_1,\ldots,\delta_m)$, $\binom{r+d-1}{r}$ denotes the number of sets of r samples, such that the minimizer obtained upon removing the samples in any of these sets violates the constraints corresponding to all r samples of that set; see Theorem 4.1 in [Gaertner & Welzl, "A Simple Sampling Lemma: Analysis and Applications in Geometric Optimization", Discrete Comput. Geom. vol. 35, pp. 569-590, 2001], and Theorem 2.3 in [Matusek, "On Geometric Optimization with Few Violated Constraints", Discrete Comput. Geom. vol. 14, pp. 365-384, 1995]. Figure 4.1 provides a pictorial illustration of this fact.

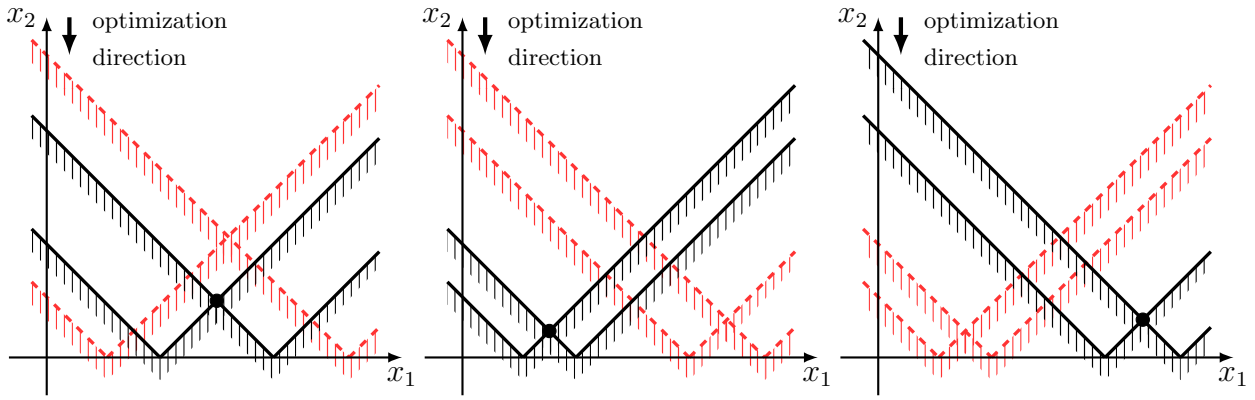

FIG. 4.1. *Pictorial example of a convex scenario program with $n_x = 2$ decision variables, $x_1, x_2 \in \mathbb{R}$, and $m = 4$ samples giving rise to V-shaped constraints. The downwards pointing arrow indicates the optimization direction, i.e., minimizing x_2, while the feasibility region is outside the shaded area. Assume that each V-shaped constraint is a translation of the other with its vertex being sampled along the x_1-axis from a distribution that admits a density, thus rendering the scenario program non-degenerate and in fact fully-supported with $\mathbb{P}^m$-probability one, i.e., $d = n_x = 2$. We discard $r = 2$ samples/constraints (the ones indicated by "dashed-red"), and the resulting solution drifts to the point indicated by the "dot". There are exactly $\binom{r+d-1}{r} = 3$ sets of $r = 2$ samples that can be discarded, such that the minimizer obtained upon the sample removal violates the constraints corresponding to the discarded samples, i.e., the associated hypothesis is not consistent with the two samples discarded each time. These three sets of discarded samples that exhibit this property are depicted in the three figure panels.*

where the first equality is due to the definition of a probability as an integral, and since for each $(\delta_1, \ldots, \delta_m)$ the summation inside the integral is equal to $\binom{r+d-1}{r}$ (by footnote 3, for each multisample this is exactly the number of nonzero terms in that summation), while the inequality is due to the fact that for any given $(\delta_1, \ldots, \delta_m)$, if $d_{\mathbb{P}}(T, H_{\bar{m}_d^r}) > \epsilon$ then $d_{\mathbb{P}}(T, H_{m_d^r}) > \epsilon$. The second last equality is by the interplay between integral and probability, and the last one is due to (3.7) specialized to the case where Assumption 4 is strengthened. By (4.1),

$$\mathbb{P}^m\Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ d_{\mathbb{P}}(T, H_{\bar{m}_d^r}) > \epsilon\Big\} \leq \sum_{i=0}^{r+d-1} \binom{m}{i} \epsilon^i (1-\epsilon)^{m-i}, \tag{4.2}$$

which is a sharper result compared to $\binom{r+d-1}{r} \sum_{i=0}^{r+d-1} \binom{m}{i} \epsilon^i (1-\epsilon)^{m-i}$. However, it holds only for the class of fully-supported, convex optimization programs, and for the constructed removal scheme that returns the solution with the lowest probability of constraint violation, while to compute the latter requires knowledge of $\mathbb{P}$ which might be unavailable. A byproduct of this fact is that $r + d$ constitutes the cardinality of a compression set for this problem, i.e., there exist $r + d$ samples such that following the same procedure using only these samples yields the same solution had all the samples been employed. The result is tight, i.e., we have equality in (4.2), if the sets of multisamples for which the probability of constraint violation for each of the $\binom{r+d-1}{r}$ solutions that violate the r removed samples exceeds ϵ, are equally likely. This occurs when we have that $\mathbb{P}^m\Big\{(\delta_1, \ldots, \delta_m) \in \Delta^m : \ H_{m_d^r}$ is not consistent with $\big\{\big(\delta_i, \mathbb{1}_T(\delta_i)\big)\big\}_{i \in I_r}$ and $d_{\mathbb{P}}(T, H_{m_d^r}) > \epsilon\Big\}$ is the same for each of the $\binom{r+d-1}{r}$ sets I_r giving rise to $H_{m_d^r}$. In that case the inequality in (4.1) becomes an equality. An example where this is the case, is the scenario program corresponding to the minimum width interval when the samples are generated in an i.i.d. fashion from a uniform distribution on $[0, 1]$.

This derivation is inspired by a dual in some sense analysis in Section 5.2 of [20], where it is established that the right-hand side of (4.2) constitutes a lower bound for the case where the hypothesis is generated by an algorithm that returns the solution with the highest probability of constraint violation instead.

5. Scenario approach and support constraints. Consider the setting of Section III-B. Helly's dimension is defined (see Definition 3.1 in [9]) as the least integer $\zeta < \infty$ such that for any finite $m \geq 1$, with $\mathbb{P}^m$-probability one with respect to the choice of an m-multisample, the number of support samples is *at most* ζ, i.e.,

$$\max_{m \geq 1} \ \text{ess sup}_{\{\delta_i\}_{i=1}^m \in \Delta^m} \Big\{\#\text{ of support constraints of } \mathcal{P}[\{\delta_i\}_{i=1}^m]\Big\} \leq \zeta. \tag{5.1}$$

It follows then that we ought to have $m \geq \zeta$. Notice that ζ is independent of the multi-sample; this implies that with non-zero probability there would exist a multisample of some length where the support constraints would be equal to ζ (otherwise a tighter bound would exist), but this is not necessarily the case for all multisamples.

PROPOSITION 5.1. *Consider Assumption 5 and fix $d = \zeta$. For any $m \geq d$, with $\mathbb{P}^m$-probability one with respect to the choice of an m-multisample $\{\delta_1, \ldots, \delta_m\}$, there exists $I_d \in \mathcal{I}_d$ such that $x_d(\{\delta_i\}_{i\in I_d}) = x_m(\{\delta_i\}_{i=1}^m)$, where $x_m(\{\delta_i\}_{i=1}^m)$ is the minimizer of $\mathcal{P}[\{\delta_i\}_{i=1}^m]$, and $x_d(\{\delta_i\}_{i\in I_d})$ denotes the optimizer of the same problem when fed only with the samples $\{\delta_i\}_{i\in I_d}$.*

Proof. We show this by means of induction.
Base case $m = d$: The statement holds trivially for any d-multisample, as we have exactly d samples.
Induction hypothesis for some $m > d$: Consider an arbitrary $m > d$. Suppose that with $\mathbb{P}^m$-probability one with respect to an m-multisample $\{\delta_1, \ldots, \delta_m\}$, there exists $I_d \in \mathcal{I}_d$ such that $x_d(\{\delta_i\}_{i\in I_d}) = x_m(\{\delta_i\}_{i=1}^m)$.
The $(m+1)$-th case: Consider an $(m+1)$-multisample $\{\delta_1, \ldots, \delta_m, \delta_{m+1}\}$. Since the number of support constraints is bounded by $d = \zeta$ (independently of the multisample length) and by the definition of ζ, with $\mathbb{P}^{m+1}$-probability one, at least $m + 1 - d > 1$ of these samples will *not* be of support. Pick one of these, and without loss of generality assume this is the $(m+1)$-th sample, namely, δ_{m+1}. By the definition of support constraints, removing the constraint associated to this sample will not change the (unique under Assumption 5) solution. We thus have that for the remaining samples $\{\delta_1, \ldots, \delta_m\}$ there exists $I_d \in \mathcal{I}_d$ such that

$$x_{m+1}(\{\delta_i\}_{i=1}^{m+1}) = x_m(\{\delta_i\}_{i=1}^m) = x_d(\{\delta_i\}_{i\in I_d}), \tag{5.2}$$

where the first equality is due to the fact that δ_{m+1} is not of support, and the second one follows from the induction hypothesis applied to the remaining m-multisample $\{\delta_1, \ldots, \delta_m\}$ (notice that the hypothesis refers to any multisample of length m). Overall, there exists a subset of the $(m+1)$-multisample with length d that results in $x_{m+1}(\{\delta_i\}_{i=1}^{m+1})$ up to a $\mathbb{P}^{m+1}$-measure zero set (follows from the fact that the induction hypothesis holds up to a $\mathbb{P}^m$-measure zero set), thus concluding the induction proof. ∎

A direct consequence of the construction in the proof, is that in the set of ζ samples that are sufficient to return the same solution with the one that would have been obtained if all the samples were used, the support samples are always included. To see this, notice that if a support sample was not included in that set, i.e., it was removed from the m-multisample, then the solution would have to change by definition of support samples/constraints. A set of ζ samples that satisfies the statement of Proposition 1 is not necessarily unique; in the particular case where the number of support constraints is *equal to* $\zeta < \infty$, i.e., (5.1) holds with equality, then such a set is unique and is equal to the set of support samples.

If the number of support constraints is *at most* $\zeta = n_x$, then Proposition 1 holds for any m-multisample as opposed to almost surely since the number of support constraints for convex problems cannot exceed n_x (e.g., see Theorem 3 in [7]), while if it is exactly *equal to* n_x we need to exclude measure zero cases that would prevent this from happening always, like selecting the same sample m times.