کنترل‌کردن سوییدگی در استفاده کردن از اطّلاعات

ارائه‌ی پروژه‌ی درس تئوری اطّلاعات

بهراد منیری محمّدرضا رحمانی

استاد درس
دکتر میرمحسنی

دستیار آموزشی
امیرحسین بساره

دانشکده‌ی مهندسی برق
دانشگاه صنعتی شریف

فهرست

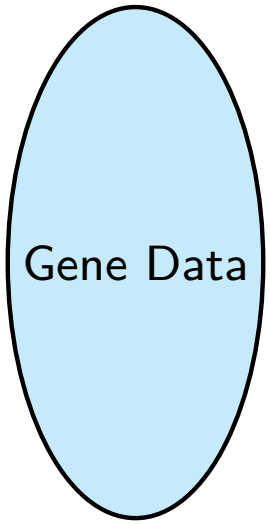
How much does your data exploration overfit? Controlling bias via information usage

Daniel Russo[1], James Zou[2]

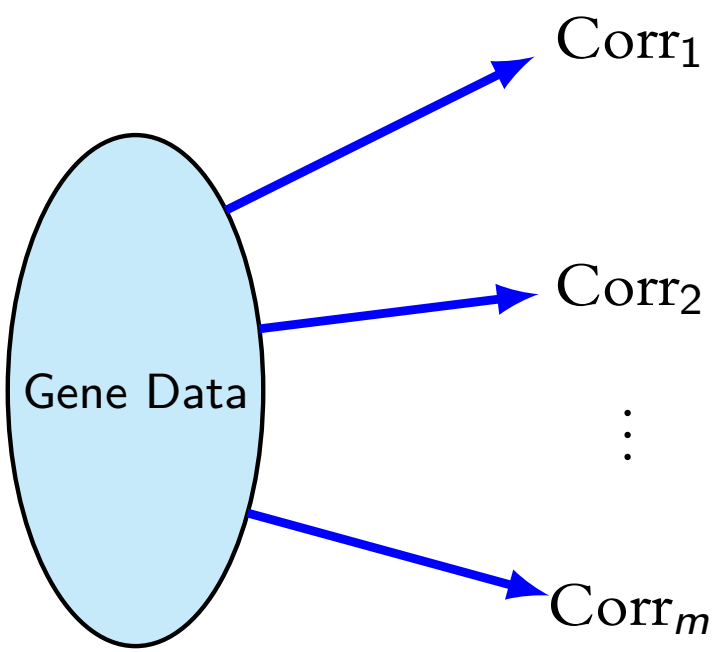
IEEE Transactions on Information Theory, 2019

[1]The Division of Decision, Risk and Operations, Columbia University
[2]Biomedical Data Science, Computer Science and Electrical Engineering at Stanford University

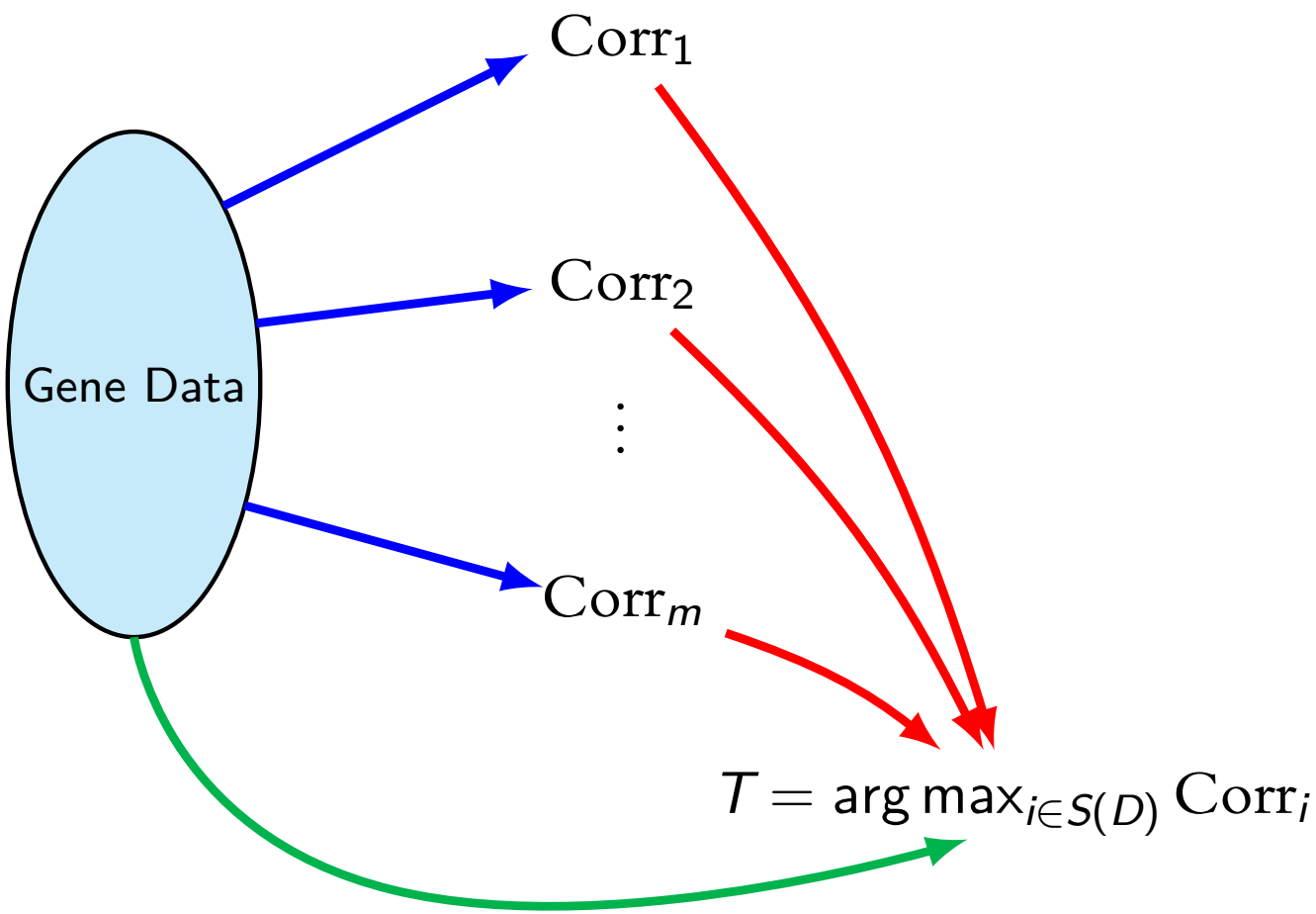
انگیزه



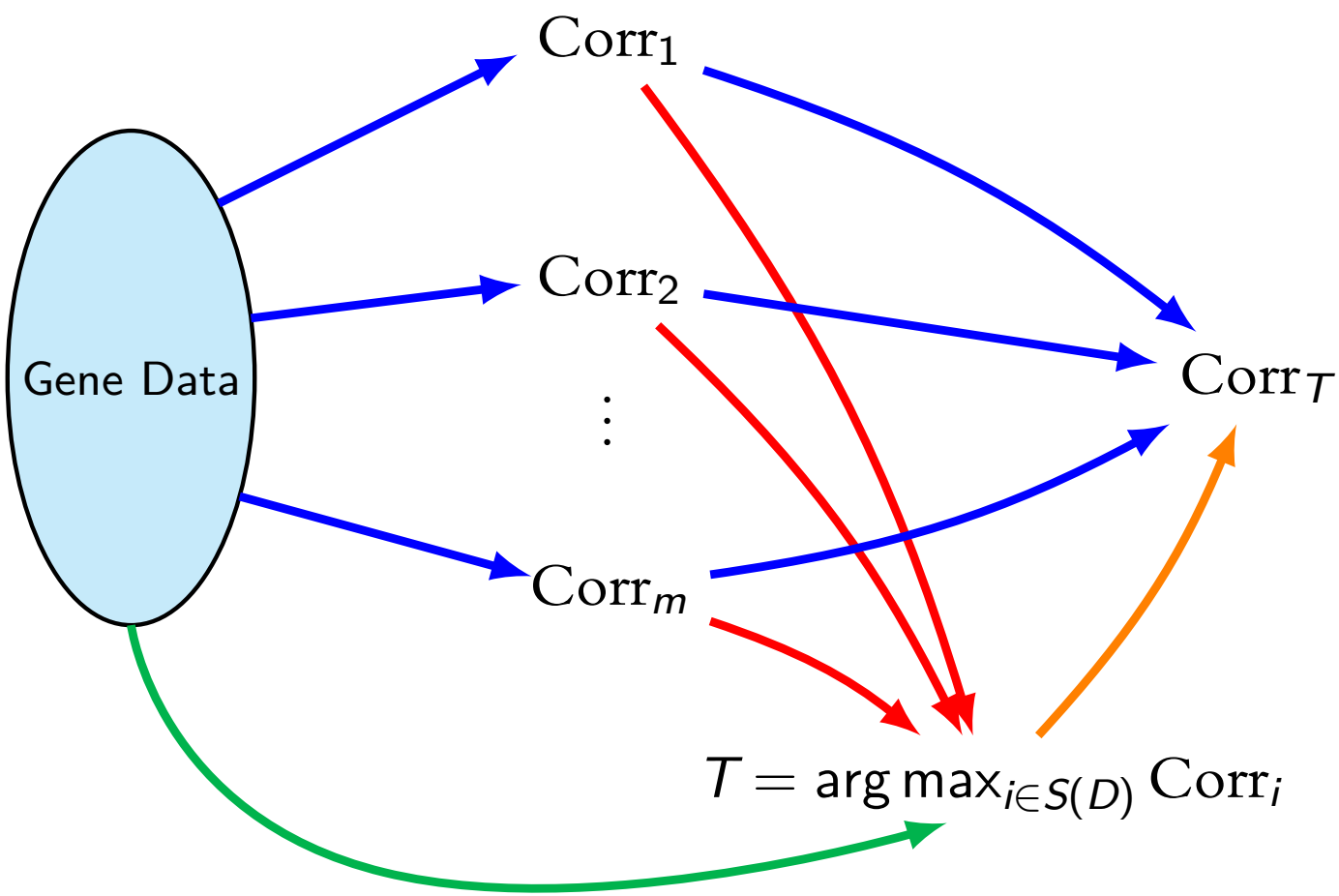
انگیزه



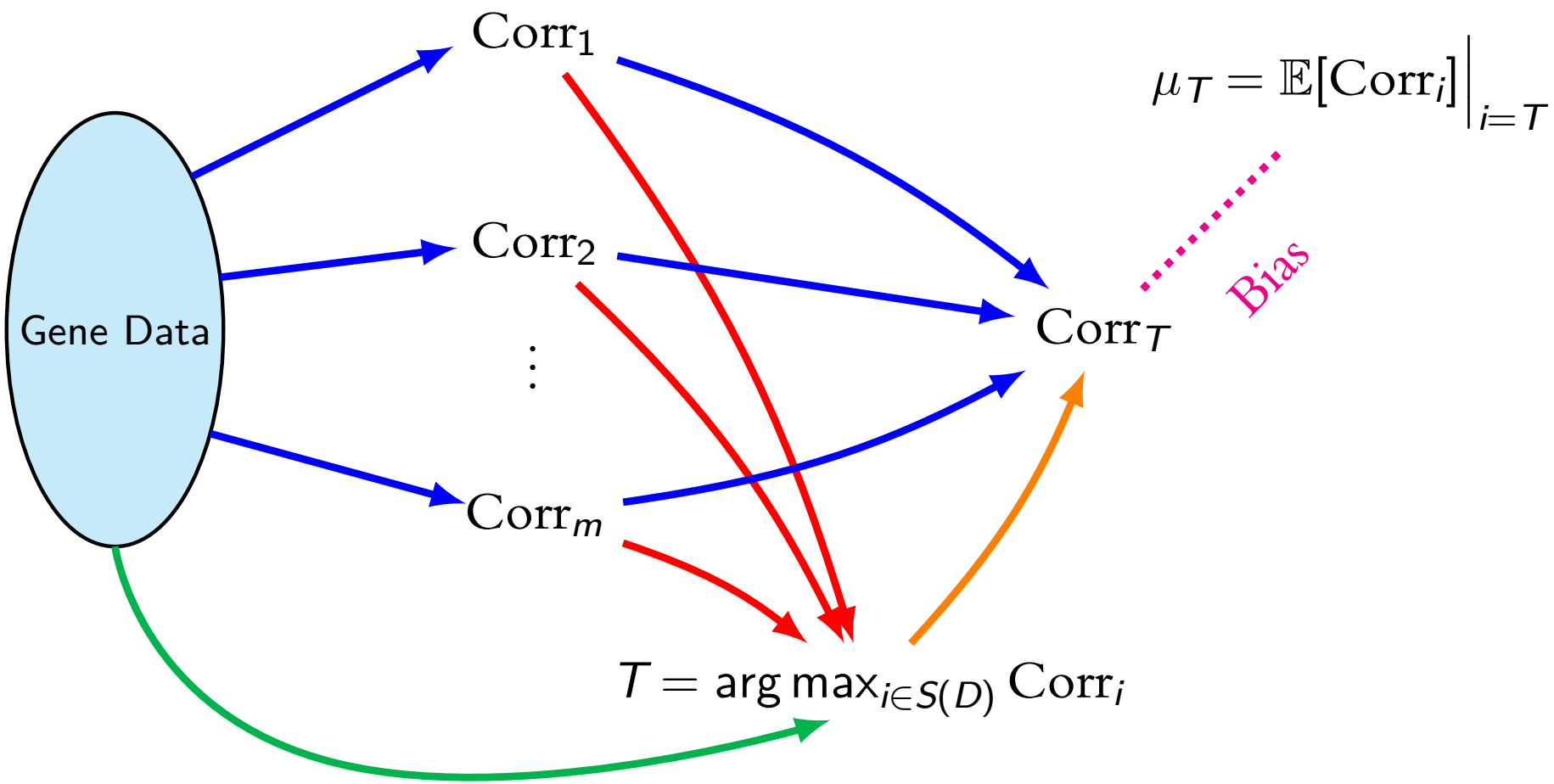
انگیزه


Gene Data
Corr_1
Corr_2
Corr_m
$T = \arg\max_{i \in S(D)} \mathrm{Corr}_i$

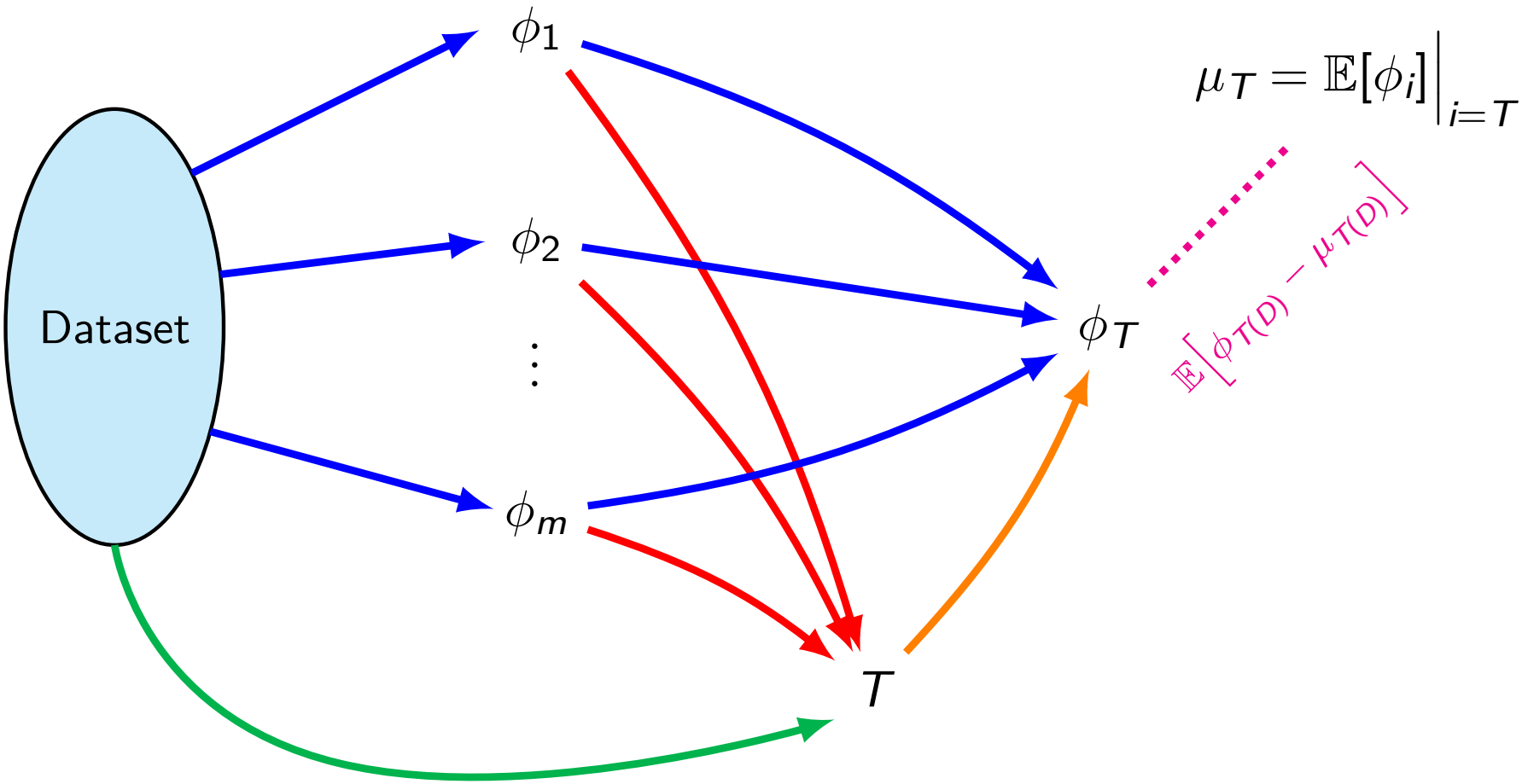
انگیزه


Gene Data
Corr_1
Corr_2
Corr_m
Corr_T
$T = \arg\max_{i \in S(D)} \mathrm{Corr}_i$

انگیزه


Gene Data
Corr_1
Corr_2
Corr_m
Corr_T
$\mu_T = \mathbb{E}[\mathrm{Corr}_i]\Big|_{i=T}$
Bias
$T = \arg\max_{i \in S(D)} \mathrm{Corr}_i$

انگیزه


Dataset
ϕ_1
ϕ_2
ϕ_m
ϕ_T
T
$\mu_T = \mathbb{E}[\phi_i]\Big|_{i=T}$
$\mathbb{E}\left[\phi_{T(D)} - \mu_{T(D)}\right]$

قضیه

متغیّر تصادفی برداری $\phi = (\phi_1, \cdots, \phi_m)$ را در نظر بگیرید، اگر تعریف کنیم

$$\mu = (\mu_1, \mu_2, \cdots, \mu_m) = \mathbb{E}[\phi]$$

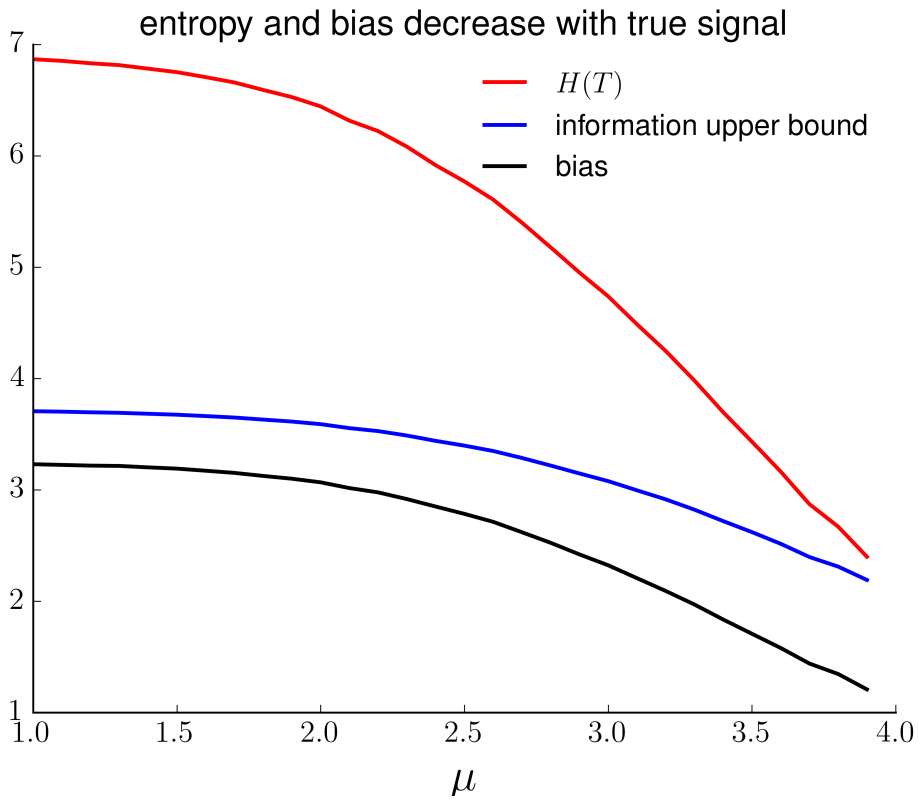
و اگر برای هر $i \in \{1, \cdots, m\}$، $\phi_i - \mu_i$ یک متغیّر تصادفی زیرگاوسی با پارامتر σ باشد، آن‌گاه:

$$\left|\mathbb{E}[\phi_T - \mu_T]\right| \leq \sigma\sqrt{2I(T;\phi)} \qquad (۱)$$

$$\mathbb{E}\left[\left|\phi_T - \mu_T\right|\right] \leq \sigma + 36\sigma\sqrt{2I(T;\phi)} \qquad (۲)$$

$$\mathbb{E}\left[(\phi_T - \mu_T)^2\right] \leq 1.25\sigma^2 + 10\sigma^2 I(T;\phi) \qquad (۳)$$

رتبه‌بندی با وجود سیگنال در داده‌ها



- مدل تولید داده‌ها:

$$\phi_i \sim \begin{cases} \mathcal{N}(\mu, \sigma^2) & i = I^* \\ \mathcal{N}(0, \sigma^2) & i \neq I^* \end{cases}$$

- انتخاب:

$$T = \arg\max_i \ \phi_i$$

- کران مبتنی بر آنتروپی:

$$I(T; \phi) = H(T) \leq \log(m)$$

تفکیک تقریباً مستقل داده‌های غیر i.i.d.

$$\underbrace{s_1 \text{ — } \cdots \text{ — } s_{n_1}}_{\text{انتخاب تحلیل}} \text{ — } s_{n_1+1} \text{ — } \cdots \text{ — } s_{n_2} \text{ — } \underbrace{s_{n_2+1} \text{ — } \cdots \text{ — } s_t}_{\text{تخمین}}$$

$$\forall \tau \in \mathbb{N} \quad \max_s D\Big(\mathbf{P}(s_\tau = .|s_1 = s)||\pi\Big) \leq c_0 e^{-c_1 \tau}$$

$$T \leftarrow \{s_1, \ldots, s_{n_1}\} - s_{n_1} - s_{n_2+1} - \{s_{n_2+1}, \ldots, s_t\} \rightarrow \phi$$

$$\begin{aligned} I(T;\phi) &\leq I(s_{n_1}; s_{n_2+1}) \\ &\leq c_0 e^{-c_1(n_2-n_1)} \end{aligned}$$

انتخاب تصادفی اندیس

- چرا تصادفی‌سازی خوب است؟

$$I(T;\phi) = H(T) - H(T|\phi) = H(T) - H(\pi)$$

- در مسئله‌ی انتخاب ماکزیمم چه می‌توان کرد؟

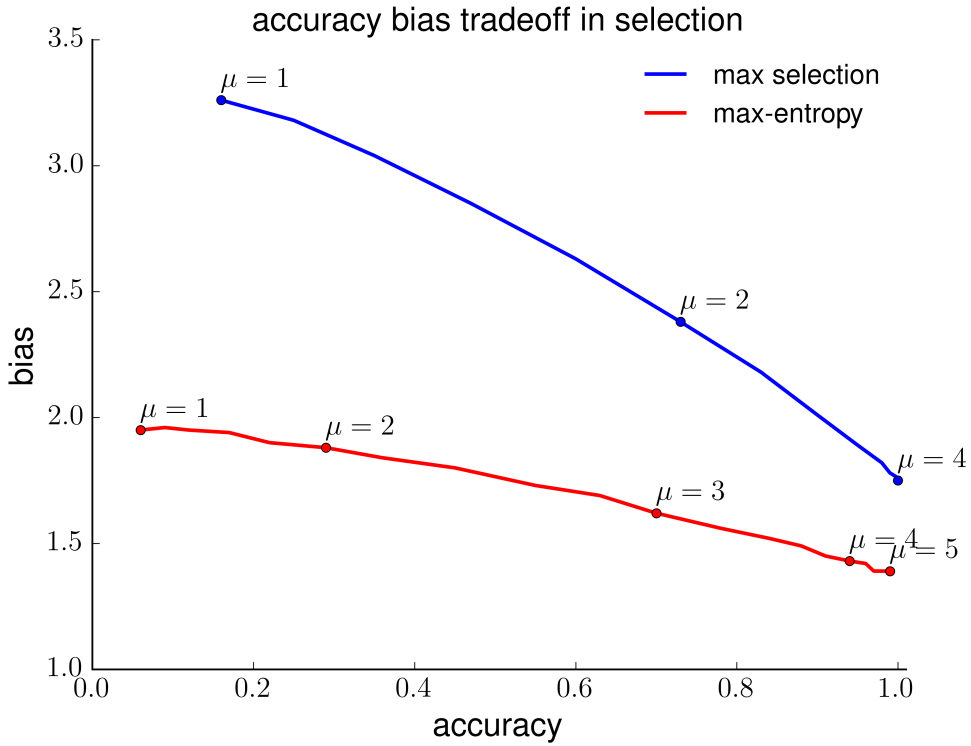
$$\underset{\pi \in \mathbb{R}_+^m}{\text{maximize}} \quad H(\pi)$$

$$\text{subject to} \quad \sum_{i=1}^{k} \pi_i \phi_i \geq b \text{ and } \sum_{i=1}^{k} \pi_i = 1.$$

پاسخ این مسئله‌ی بهینه‌سازی:

$$\pi^* = Ae^{\beta \phi_i}$$

انتخاب تصادفی اندیس



- نحوه‌ی تولید داده‌ها:

$$\begin{cases} \mu_i = \mu & i \leq N_1 \\ \mu_i = 0 & N_1 \leq i \leq N \end{cases}$$

سوییدگی:

$$\frac{1}{K}\sum_{i=1}^{K} \phi_{T_i} - \mu_{T_i}$$

دقّت:

$$\frac{\left|\{T_i : T_i \leq N_1\}\right|}{K}$$

مدل تحلیل تطبیقی داده‌ها

- در قدم اول، تحلیل ϕ_{T_1} انتخاب می‌شود. نتیجه‌ی این تحلیل $Y_{T_1} \in \mathbb{R}$ است.
- در تکرار k ام:

$$T_k = f\big(Y_{T_1}, Y_{T_2}, \ldots, Y_{T_k-1}, T_1, \ldots, T_{k-1}\big) \Rightarrow \phi_{T_k} \Rightarrow Y_{T_k}$$

گزاره

سوییدگی Y_{T_k} دو جزء دارد:

$$\begin{aligned}
&\mathbb{E}[|Y_{T_k} - \mu_{T_k}|] - \sigma \\
&\leq \mathbb{E}[|Y_{T_k} - \phi_{T_k}|] + \mathbb{E}[|\phi_{T_k} - \mu_{T_k}| - \sigma] \\
&\leq \underbrace{\mathbb{E}[|Y_{T_k} - \phi_{T_k}|]}_{\text{اعوجاج}} + \underbrace{c\sigma\sqrt{2I(T_k;\phi)}}_{\text{سوییدگی ناشی از انتخاب}}.
\end{aligned}$$

مدل تحلیل تطبیقی داده‌ها

$$T_{k+1} \leftarrow H_k = \{T_1, Y_{T_1}, T_2, Y_{T_2}, \ldots, T_k, Y_{T_k}\} \leftarrow D \rightarrow \phi$$

لم

$$I(T_{k+1}; \phi) \leq I(H_k; \phi) = \sum_{i=1}^{k} I(Y_{T_i}; \phi_{T_i} | H_{i-1}, T_i)$$

مدل تحلیل تطبیقی داده‌ها

$$T_{k+1} \leftarrow H_k = \{T_1, Y_{T_1}, T_2, Y_{T_2}, \ldots, T_k, Y_{T_k}\} \leftarrow D \rightarrow \phi$$

لم

$$I(T_{k+1}; \phi) \leq I(H_k; \phi) = \sum_{i=1}^{k} I(Y_{T_i}; \phi_{T_i} | H_{i-1}, T_i)$$

ایده

از آن‌جایی که باید اطّلاعات متقابل Y_{T_k} و ϕ_{T_k} را کنترل کنیم:

$$Y_{T_i} = \phi_{T_i} + \text{نویز}$$

پیش‌نهادها

- انتخاب تعدادی تصادفی از تحلیل‌ها به جای تنها یک تحلیل
- یافتن توزیع نویز بهینه در مسئله‌ی تحلیل تطبیقی برای کمترین سوییدگی
- بررسی الگوریتم‌های آنلاین دیگر در چهارچوب اطلاعات استفاده‌شده
- مدل‌کردن دیتاست به صورت یک فرآیند تصادفی

مراجع

Russo, D. and J. Zou (2016).

Controlling bias in adaptive data analysis using information theory.

In *Artificial Intelligence and Statistics*, pp. 1232–1240.

Russo, D. and J. Zou (2019).

How much does your data exploration overfit? controlling bias via information usage.

IEEE Transactions on Information Theory 66(1), 302–323.

Wainwright, M. J. (2019).

High-dimensional statistics: A non-asymptotic viewpoint, Volume 48.

Cambridge University Press.

با تشکّر