Variance reduction techniques for importance-weighted cross-validation.

ICT.OPEN 21-03-2017
WM Kouw & M Loog
Pattern Recognition Lab

Covariate shift

- **In domain adaptation, one hopes to generalize from a *source* domain to a *target* domain.**

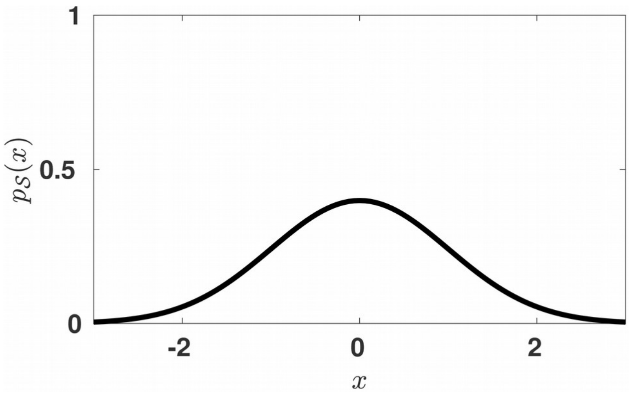
Covariate shift

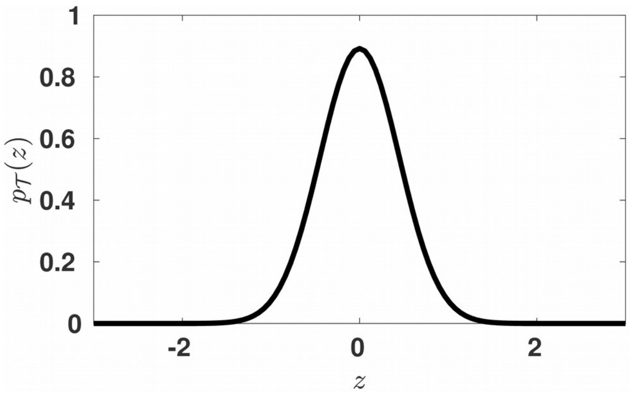
- **In domain adaptation, one hopes to generalize from a *source* domain to a *target* domain.**
  - Domains are different probability measures over the same sample space.

Covariate shift

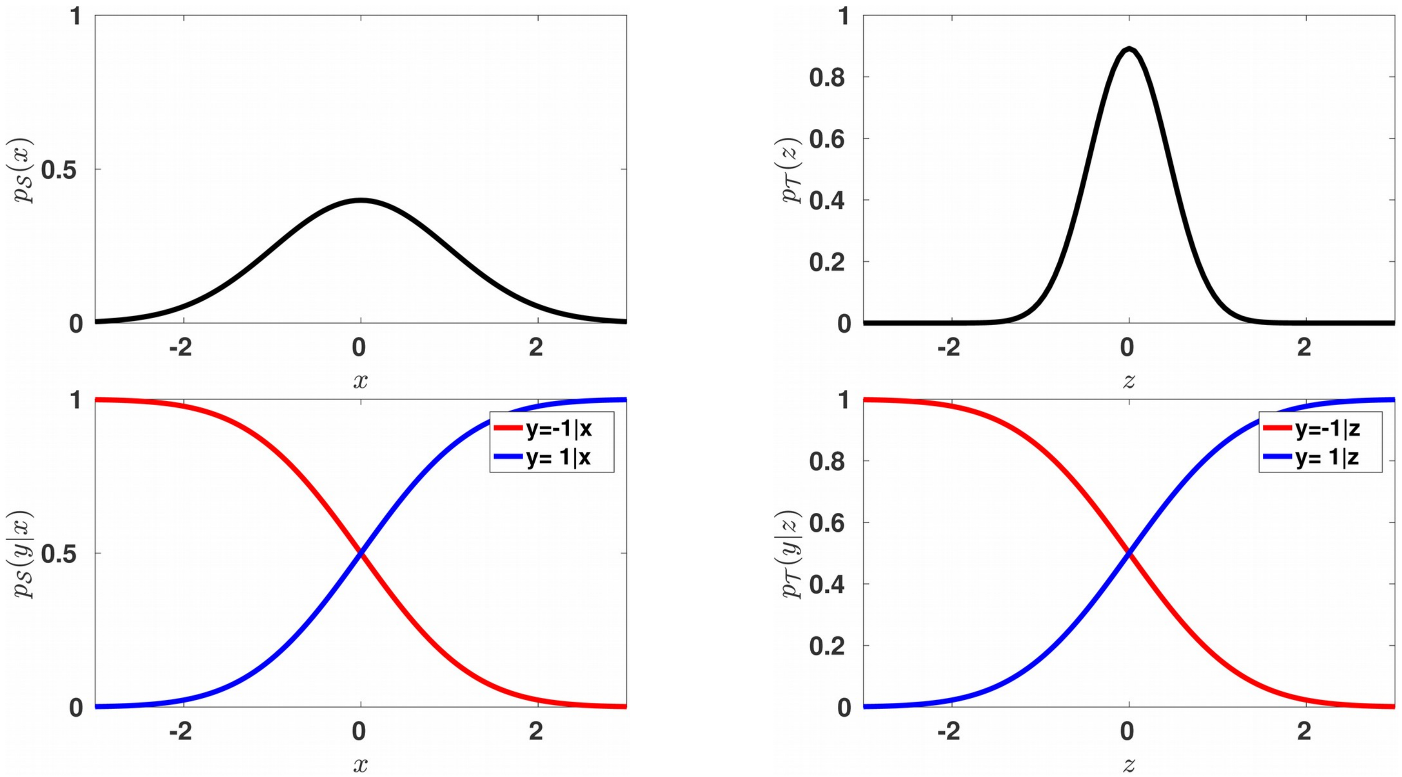
- **In domain adaptation, one hopes to generalize from a *source* domain to a *target* domain.**
  - Domains are different probability measures over the same sample space.
- **Covariate shift is the particular case where the class-posterior distributions are equivalent:** $p_{\mathcal{S}}(y \mid x) = p_{\mathcal{T}}(y \mid x)$

Covariate shift

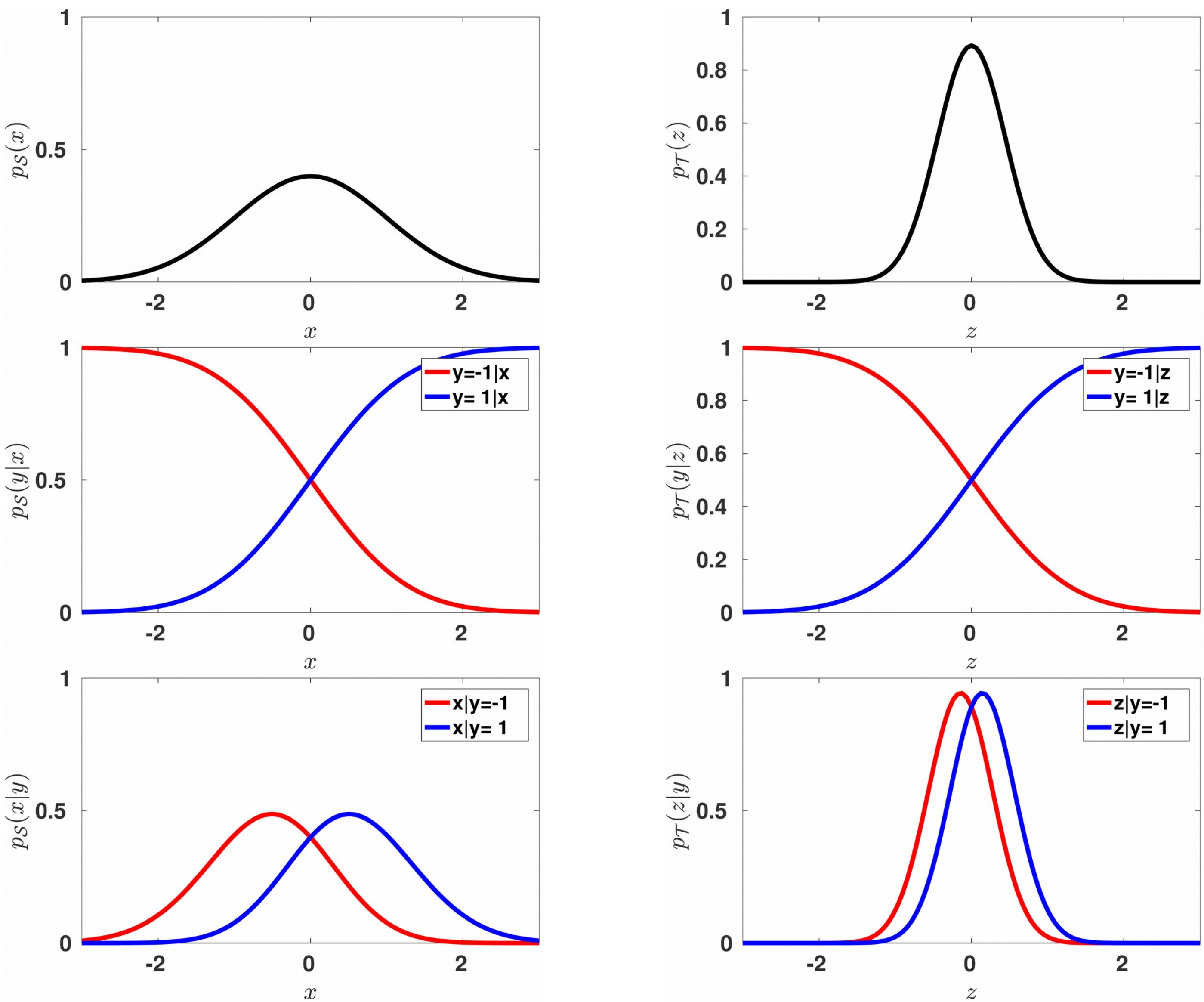




Covariate shift



Covariate shift



Covariate shift

- **One can rewrite the risk functional as follows:**

$$\int_{\Omega} \sum_{y} \ell(h(x), y) p_{\mathcal{T}}(x, y) \, \mathrm{d}x = \int_{\Omega} \sum_{y} \ell(h(x), y) \frac{p_{\mathcal{T}}(x, y)}{p_{\mathcal{S}}(x, y)} p_{\mathcal{S}}(x, y) \, \mathrm{d}x$$

Covariate shift

- **One can rewrite the risk functional as follows:**

$$\int_\Omega \sum_y \ell(h(x), y) p_\mathcal{T}(x, y) \; \mathrm{d}x = \int_\Omega \sum_y \ell(h(x), y) \frac{p_\mathcal{T}(x, y)}{p_\mathcal{S}(x, y)} p_\mathcal{S}(x, y) \; \mathrm{d}x$$

- **Setting the class-posterior distributions equal leads to:**

$$\int_\Omega \sum_y \ell(h(x), y) p_\mathcal{T}(x, y) \; \mathrm{d}x = \int_\Omega \sum_y \ell(h(x), y) \frac{p_\mathcal{T}(x)}{p_\mathcal{S}(x)} p_\mathcal{S}(x, y) \; \mathrm{d}x$$

Covariate shift

- **One can rewrite the risk functional as follows:**

$$\int_\Omega \sum_y \ell(h(x),y) p_\mathcal{T}(x,y) \ \mathrm{d}x = \int_\Omega \sum_y \ell(h(x),y) \frac{p_\mathcal{T}(x,y)}{p_\mathcal{S}(x,y)} p_\mathcal{S}(x,y) \ \mathrm{d}x$$

- **Setting the class-posterior distributions equal leads to:**

$$\int_\Omega \sum_y \ell(h(x),y) p_\mathcal{T}(x,y) \ \mathrm{d}x = \int_\Omega \sum_y \ell(h(x),y) \frac{p_\mathcal{T}(x)}{p_\mathcal{S}(x)} p_\mathcal{S}(x,y) \ \mathrm{d}x$$

- **In covariate shift, the target risk is equivalent to the weighted source risk.**

Importance weighing

- Weighing can be viewed as importance sampling the source domain in the area of the target domain.

Importance weighing

- **Weighing can be viewed as importance sampling the source domain in the area of the target domain.**
  - Original target risk estimator is: $\{(x,y)\}_{i=1}^{n} \sim p_{\mathcal{S}}(x,y)$

Importance weighing

- **Weighing can be viewed as importance sampling the source domain in the area of the target domain.**
  - Original target risk estimator is: $$\hat{R}_{\mathcal{T}} = \frac{1}{m} \sum_{j=1}^{m} \ell(h(z_j|\theta), u_j)$$

Importance weighing

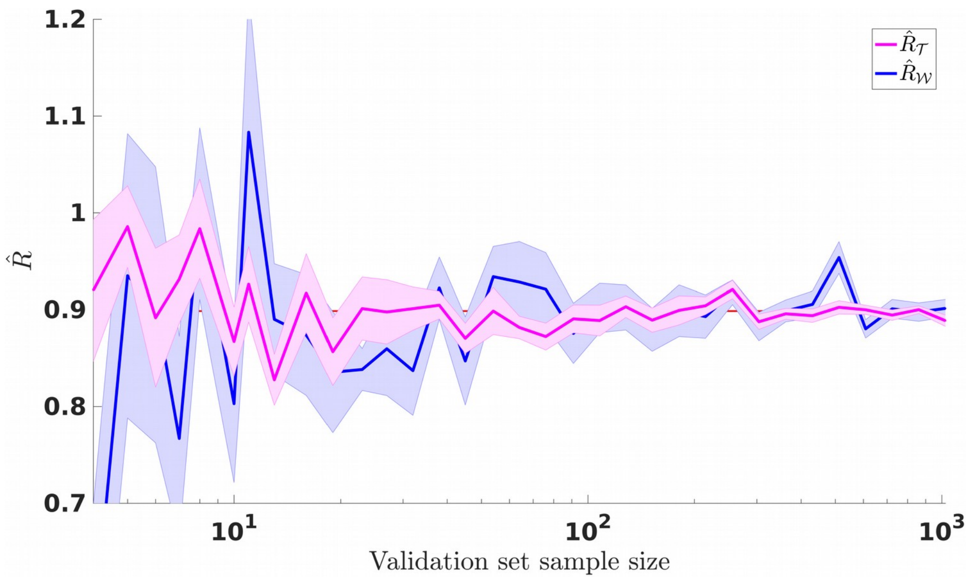
- **Weighing can be viewed as importance sampling the source domain in the area of the target domain.**
  - Original target risk estimator is: $\hat{R}_{\mathcal{T}} = \frac{1}{m}\sum_{j=1}^{m} \ell(h(z_j|\theta), u_j)$
  - Importance-weighted source risk is: $\{(z,u)\}_{j=1}^{m} \sim p_{\mathcal{T}}(x,y)$

Importance weighing

- **Weighing can be viewed as importance sampling the source domain in the area of the target domain.**
  - Original target risk estimator is: $$\hat{R}_{\mathcal{T}} = \frac{1}{m} \sum_{j=1}^{m} \ell(h(z_j|\theta), u_j)$$
  - Importance-weighted source risk is: $$\hat{R}_{\mathcal{W}} = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i|\theta), y_i) \frac{p_{\mathcal{T}}(x_i)}{p_{\mathcal{S}}(x_i)}$$

Importance weighing

- **Weighing can be viewed as importance sampling the source domain in the area of the target domain.**
  - Original target risk estimator is: $$\hat{R}_{\mathcal{T}} = \frac{1}{m} \sum_{j=1}^{m} \ell(h(z_j|\theta), u_j)$$
  - Importance-weighted source risk is: $$\hat{R}_{\mathcal{W}} = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i|\theta), y_i) \frac{p_{\mathcal{T}}(x_i)}{p_{\mathcal{S}}(x_i)}$$



Sampling variance

- **However, the weights increase sampling variance:**
  - Original target risk sampling variance:

$$\mathbb{V}[\hat{R}_{\mathcal{T}}] = \frac{1}{m} \int_{\Omega} \sum_{y \in Y} \big(\ell(h(x|\theta), y) - R_{\mathcal{T}} \big)^2 p_{\mathcal{T}}(x, y) \, \mathrm{d}x$$

Sampling variance

- **However, the weights increase sampling variance:**
  - Original target risk sampling variance:

$$\mathbb{V}[\hat{R}_{\mathcal{T}}] = \frac{1}{m} \int_{\Omega} \sum_{y \in Y} \big(\ell(h(x|\theta), y) - R_{\mathcal{T}}\big)^2 p_{\mathcal{T}}(x, y) \ \mathrm{d}x$$

  - Importance-weighted risk sampling variance:

$$\mathbb{V}[\hat{R}_{\mathcal{W}}] = \frac{1}{n} \int_{\Omega} \sum_{y \in Y} \big(\ell(h(x|\theta), y) \frac{p_{\mathcal{T}}(x)}{p_{\mathcal{S}}(x)} - R_{\mathcal{T}}\big)^2 p_{\mathcal{S}}(x, y) \ \mathrm{d}x$$

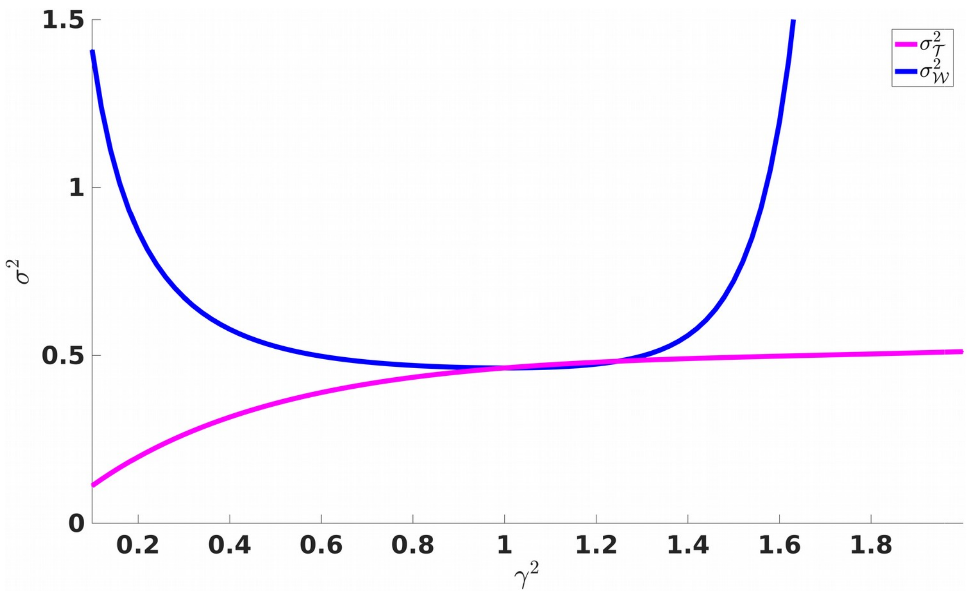
Sampling variance

- **However, the weights increase sampling variance:**
  - Original target risk sampling variance:

$$\mathbb{V}[\hat{R}_{\mathcal{T}}] = \frac{1}{m} \int_{\Omega} \sum_{y \in Y} \big(\ell(h(x|\theta), y) - R_{\mathcal{T}}\big)^2 p_{\mathcal{T}}(x, y) \ \mathrm{d}x$$

  - Importance-weighted risk sampling variance:

$$\mathbb{V}[\hat{R}_{\mathcal{W}}] = \frac{1}{n} \int_{\Omega} \sum_{y \in Y} \big(\ell(h(x|\theta), y) \frac{p_{\mathcal{T}}(x)}{p_{\mathcal{S}}(x)} - R_{\mathcal{T}}\big)^2 p_{\mathcal{S}}(x, y) \ \mathrm{d}x$$



Control variate

- If one has additional knowledge on the estimand, one can design a classifier with less sampling variance.

Control variate

- **If one has additional knowledge on the estimand, one can design a classifier with less sampling variance.**
  - Covariate shift setting: expected value of the weights is 1.

Control variate

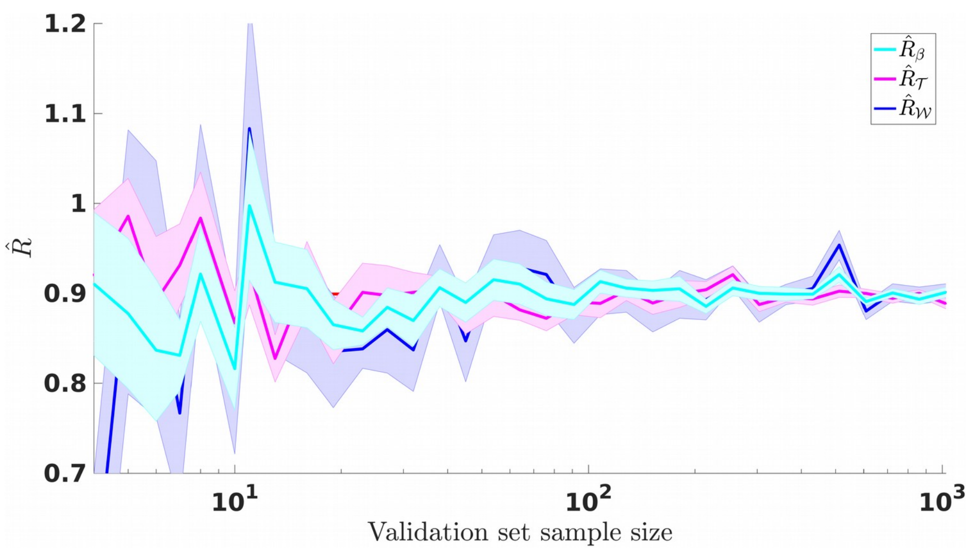
- **If one has additional knowledge on the estimand, one can design a classifier with less sampling variance.**
  - Covariate shift setting: expected value of the weights is 1.
- **Using the weights as a *control variate* leads to following estimator:**

$$\hat{R}_\beta = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i|\theta), y_i) \frac{p_\mathcal{T}(x_i)}{p_\mathcal{S}(x_i)} - \beta(\frac{p_\mathcal{T}(x_i)}{p_\mathcal{S}(x_i)} - 1)$$

Control variate

- **If one has additional knowledge on the estimand, one can design a classifier with less sampling variance.**
  - Covariate shift setting: expected value of the weights is 1.
- **Using the weights as a *control variate* leads to following estimator:**

$$\hat{R}_\beta = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i|\theta), y_i) \frac{p_\mathcal{T}(x_i)}{p_\mathcal{S}(x_i)} - \beta\left(\frac{p_\mathcal{T}(x_i)}{p_\mathcal{S}(x_i)} - 1\right)$$



Variance reduction

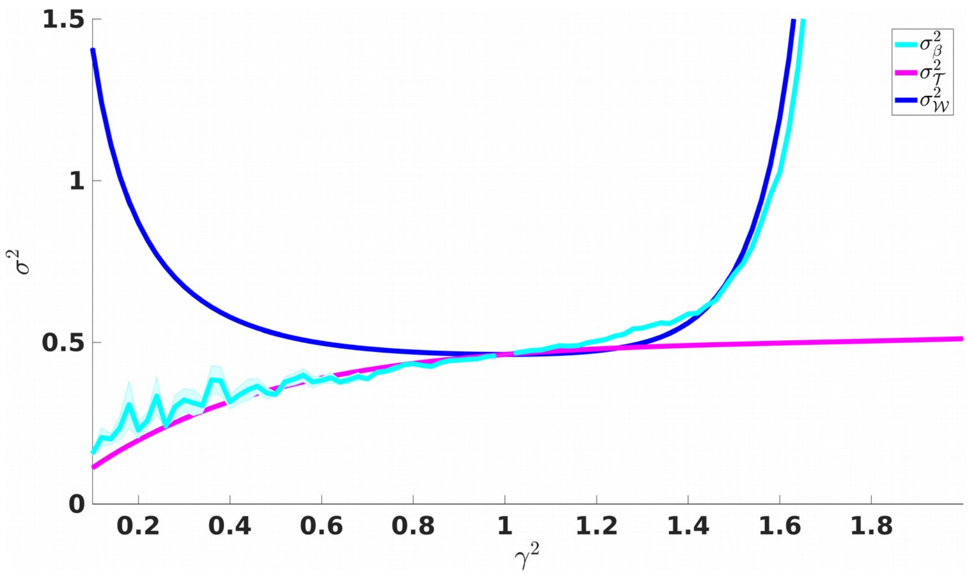
– **The control-variate-based estimator has, for optimal β, a sampling variance of:**

$$\mathbb{V}[\hat{R}_\beta] = \frac{1}{n} \int_\Omega \sum_{y \in Y} \big(\ell(h(x|\theta), y) \frac{p_\mathcal{T}(x_i)}{p_\mathcal{S}(x_i)} - \beta(\frac{p_\mathcal{T}(x_i)}{p_\mathcal{S}(x_i)} - 1) - R_\mathcal{T}\big)^2 p_\mathcal{S}(x, y) \mathrm{d}x$$

Variance reduction

– **The control-variate-based estimator has, for optimal β, a sampling variance of:**

$$\mathbb{V}[\hat{R}_\beta] = \frac{1}{n}\int_\Omega \sum_{y \in Y} \Big(\ell(h(x|\theta), y)\frac{p_\mathcal{T}(x_i)}{p_\mathcal{S}(x_i)} - \beta(\frac{p_\mathcal{T}(x_i)}{p_\mathcal{S}(x_i)} - 1) - R_\mathcal{T}\Big)^2 p_\mathcal{S}(x,y)\mathrm{d}x$$



Cross-validation

- **For the same sample size, the regression importance-weighted risk is a better estimator.**

Cross-validation

- **For the same sample size, the regression importance-weighted risk is a better estimator.**
  - A better risk estimator leads to better cross-validation and, in turn, better hyperparameters.

Questions

- **If you are interested, I'm happy to elaborate on and discuss our approaches at my poster.**