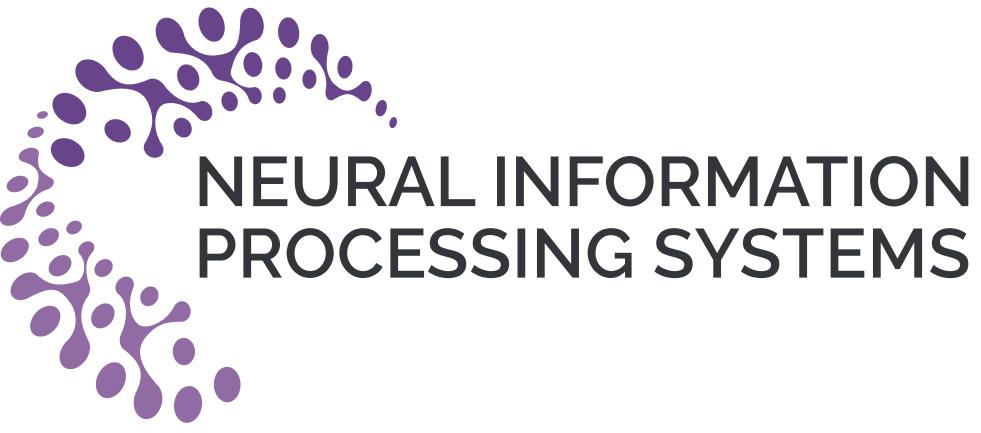
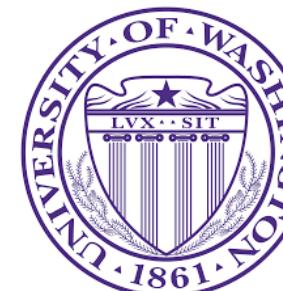


Stochastic Gradients under Nuisances

Facheng Yu, Ronak Mehta, Alex Luedke, Zaid Harchaoui



Orthogonal Statistical Learning

Motivation: Many machine learning problems rely on a risk function that is only partially specified up to a class of risks under the unknown nuisance, wherein the risk of interest is the one under the true nuisance.

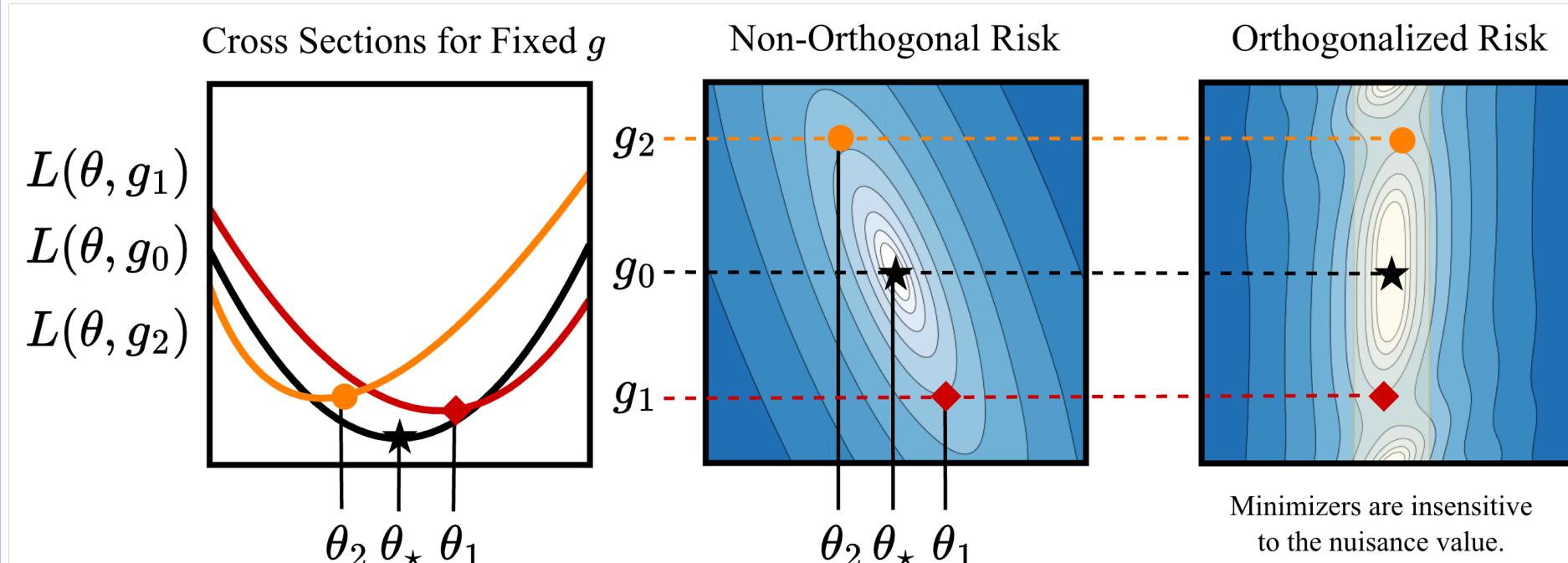
Risk function under unknown nuisance

$$\mathcal{L} := \{L(\cdot, g) : g \in \mathcal{G}\}$$

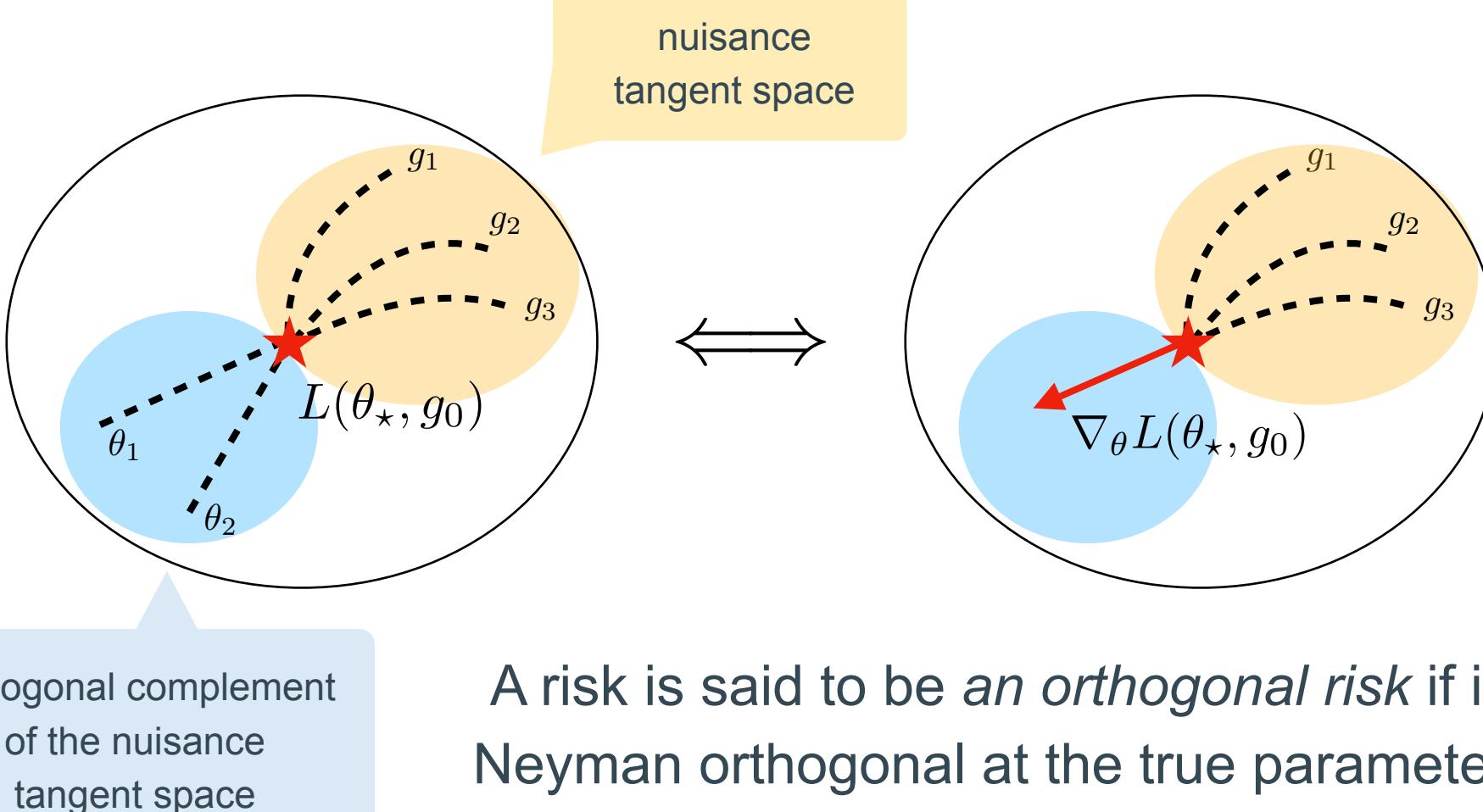
Risk minimization under true nuisance

$$\theta_* = \arg \min_{\theta \in \Theta} [L_0(\theta) := \mathbb{E}_{Z \sim \mathbb{P}} [\ell(\theta, g_0; Z)]]$$

Learning the true nuisance would introduce additional statistical error. One way to make the risk insensitive to nuisance is to **orthogonalize it**.



Intuition of Neyman Orthogonality: The parameter tangent space is **orthogonal** to the nuisance tangent space.



A risk is said to be *an orthogonal risk* if it is Neyman orthogonal at the true parameter of interest and the true nuisance.

Classical Stochastic Gradient Algorithm

$$Z_1, \dots, Z_n \stackrel{i.i.d.}{\sim} \mathbb{P}$$

$$(SGD) \quad \theta^{(n)} = \theta^{(n-1)} - \eta \nabla_{\theta} \ell(\theta^{(n-1)}, \hat{g}; Z_n)$$

learning rate
plug-in nuisance estimator
(independent of samples)

Theorem 1. The standard SGD iterate expected errors satisfy

$$O((1 - \mu\eta/2)^n + \eta + \|\hat{g} - g_0\|_{\mathcal{G}}^2) \quad (\text{Nuisance sensitive})$$

For an orthogonal risk, the iterates expected errors satisfy

$$O((1 - \mu\eta/2)^n + \eta + \|\hat{g} - g_0\|_{\mathcal{G}}^4) \quad (\text{Nuisance Insensitive})$$

SGD Optimization Error

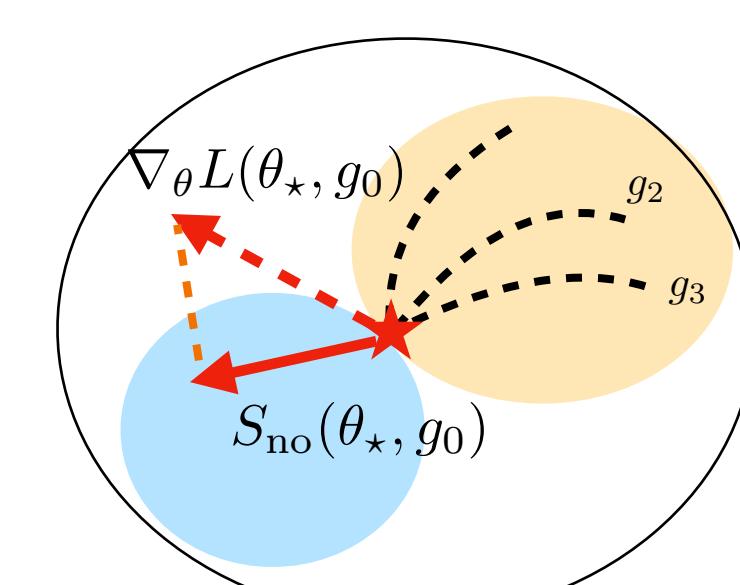
Nuisance Estimation Error

Using an orthogonal loss can further remove the nuisance estimation error. However, it is not always possible to handcraft orthogonalized objectives. Sequentially orthogonalizing first-order information is more **flexible**.

Orthogonalized Stochastic Gradients

$$S_{no}(\theta, g; z) = \nabla_{\theta} \ell(\theta, g; z) - \Gamma_0 \nabla_g \ell(\theta, g; z)$$

orthogonalizing operator



Intuition:

Project the standard gradient oracle onto the *orthogonal complement* of the nuisance tangent space.

Example. [Partially Linear Model]

$$Y = \langle \theta_*, X \rangle + g_0(W) + \epsilon \quad Z = (X, Y, W) \sim \mathbb{P}$$

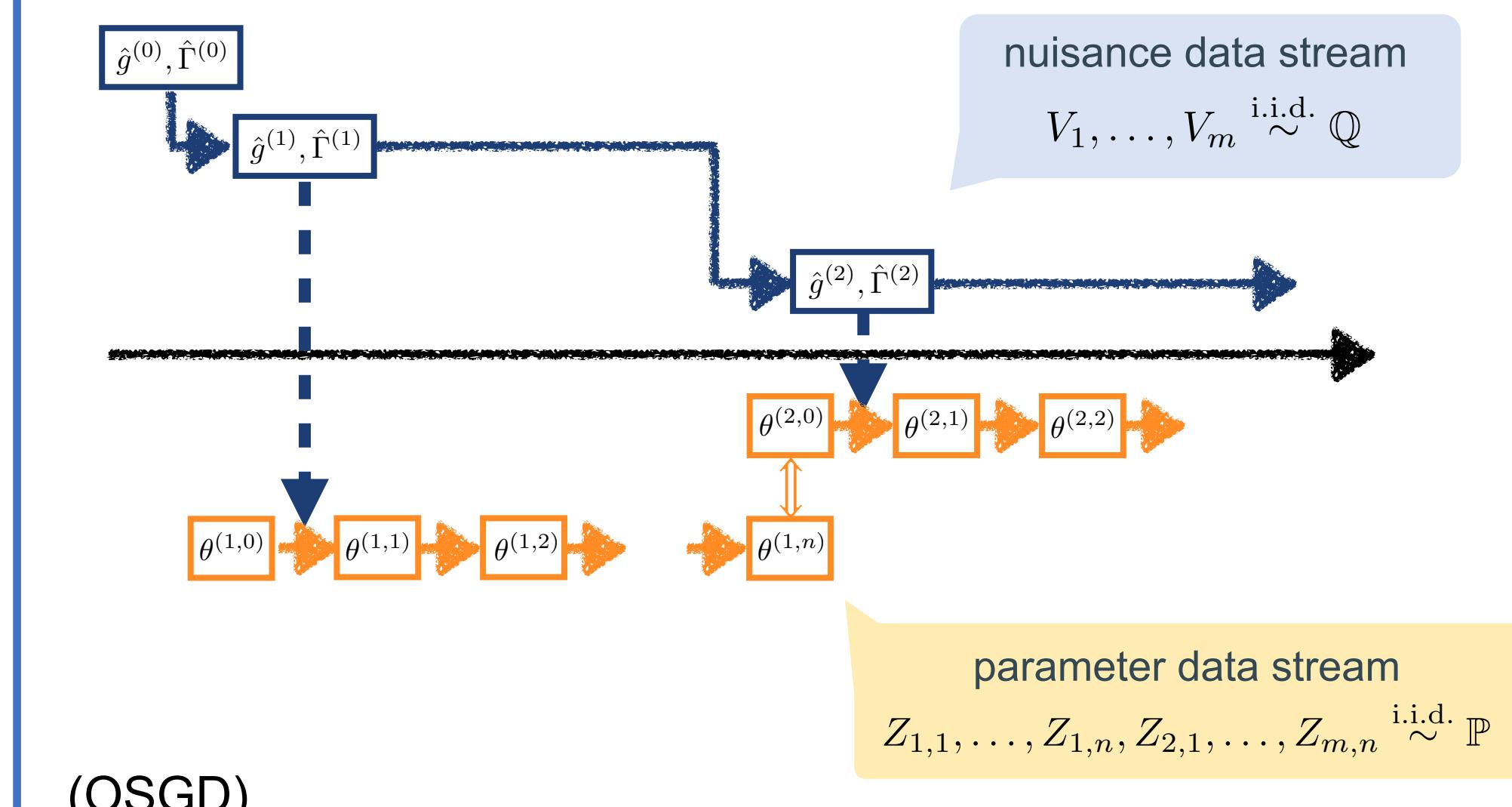
$$U = g_0(W) + \xi \quad V = (X, U, W) \sim \mathbb{Q}$$

Consider the following non-orthogonal loss

$$\ell(\theta, g; z) = \frac{1}{2}(y - g(w) - \langle \theta, x \rangle)^2$$

The true nuisance and the orthogonalizing operator are

$$g_0(W) = \mathbb{E}[U | W] \quad \Gamma_0 : g \mapsto \mathbb{E}[\mathbb{E}[X | W]g(W)]$$



$$\theta^{(m,n)} = \theta^{(n-1)} - \eta \hat{S}_{no}^{(m)}(\theta^{(m,n-1)}, \hat{g}^{(m)}; Z_{m,n})$$

approximated stochastic gradient oracle
 $\hat{S}_{no}^{(m)}(\theta, g; z) = \nabla_{\theta} \ell(\theta, g; z) - \hat{\Gamma}^{(m)} \nabla_g \ell(\theta, g; z)$

Theorem 2. OSGD iterate expected errors satisfy

$$O((1 - \mu\eta/2)^{mn} + n^{-1} + \eta + m^{-(2\alpha-1)/\alpha})$$

nuisance insensitive rate

$$\begin{aligned} \|\hat{g}^{(m)} - g_0\|_{\mathcal{G}}^2 &= O_p(m^{-(2\alpha-1)/2\alpha}) \\ \|\hat{\Gamma}^{(m)} - \Gamma_0\|^2 &= O_p(m^{-(2\alpha-1)/2\alpha}) \\ (\eta n)^{-1} &= O(1) \end{aligned}$$

