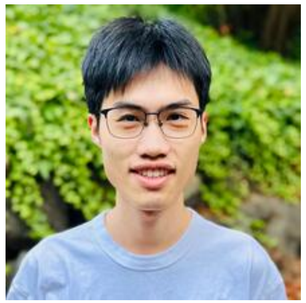
Stochastic Gradients under Nuisances

Institute for Foundations of Data Science (IFDS) Seminar
October 10, 2025

UNIVERSITY of WASHINGTON

Facheng Yu

Team



Facheng Yu
University of Washington

Ronak Mehta
University of Washington

Alex Luedtke
Harvard Medical School

Zaid Harchaoui
University of Washington

Institute for Foundations of Data Science

IFDS

Washington · Wisconsin · Santa Cruz · Chicago

UNIVERSITY of WASHINGTON

What Are Nuisances?

Nuisances: any function class $\mathcal{G} := \{g : \mathcal{Z} \mapsto \mathbb{R}\}$.

True nuisance: $g_0 \in \mathcal{G}$, which defines the learning problem

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

Usually the true nuisance is in the form of $\mathbb{E}[Y \mid X]$.

Nuisance: A perspective of regression

For a model class $\mathcal{F} = \{f_\theta : \mathcal{X} \mapsto \mathbb{R} \mid \theta \in \Theta\}$:

(Regression problem) $Y = f_{\theta_\star}(X) + \epsilon$

Residual restricted by zero-mean (conditionally), i.e., $\mathbb{E}[\epsilon \mid X] = 0$.

What would happen if an additional feature W is observed?

ϵ needs to be further explained by W as $g_0(W) + \epsilon'$ such that

$$\mathbb{E}[\epsilon' \mid X, W] = 0.$$

Example 1: Partially Linear Model

$$g_0(w) = \mathbb{E}\big[Y - \langle \theta_0, X \rangle \;\big|\; W = w\big].$$

Let $Z = (X, W, Y) \sim \mathbb{P}$ such that

$$Y = \langle \theta_0, X \rangle + g_0(W) + \epsilon, \;\; \text{where } \mathbb{E}[\epsilon \mid X, W] = 0.$$

We can estimate θ_0 via the optimization problem:

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

$$\ell(\theta, g_0; Z) = (Y - g_0(W) - \langle \theta, X \rangle)^2.$$

Scenarios with Nuisances:

$$\max_{\theta \in \Theta} \max_{g \in \mathcal{G}} \log P(X \mid \theta, g)$$

The profile g is a nuisance!

- Statistics: Semiparametric Inference
- Machine Learning: Profile Likelihood, Zero-Shot Prediction
- Optimization: Distributionally Robust Optimization

$$\min_{w \in \mathbb{R}^d} \mathcal{R}_{\mathcal{P}}(\ell(w)) \quad \text{where} \quad \mathcal{R}_{\mathcal{P}}(l) := \max_{q \in \mathcal{P}} \left\{ \sum_{i=1}^{n} q_i l_i - \nu D(q \| \mathbf{1}_n / n) \right\},$$

(R. Mehta et. al. ICML 2024)

The probability density q is a nuisance!

Nuisances: any function class $\mathcal{G} := \{g : \mathcal{Z} \mapsto \mathbb{R}\}$.

True nuisance: $g_0 \in \mathcal{G}$, which defines the learning problem

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

The true nuisance is always unknown — how to learn $\theta_\star$?

— Double Machine Learning (DML):

1. Estimate g_0 by $\hat{g}$.

2. Estimate $\theta_\star$ by $\arg\min_{\theta \in \Theta} \mathbb{E}_{Z \sim \mathbb{P}}[\ell(\theta, \hat{g}; Z)]$.

Double/Debiased Machine Learning (DML)

Double/debiased machine learning for treatment and structural parameters

VICTOR CHERNOZHUKOV[†], DENIS CHETVERIKOV[‡], MERT DEMIRER[†], ESTHER DUFLO[†], CHRISTIAN HANSEN[§], WHITNEY NEWEY[†] AND JAMES ROBINS[‖]

RieszNet and ForestRiesz: Automatic Debiased Machine Learning with Neural Nets and Random Forests

Victor Chernozhukov [1] **Whitney Newey** [1] **Víctor Quintas-Martínez** [1] **Vasilis Syrgkanis** [2]

Adapting Neural Networks for the Estimation of Treatment Effects

Claudia Shi[1], David M. Blei[1,2], and Victor Veitch[2]

[1]*Department of Computer Science, Columbia Unitversity*
[2]*Department of Statistics, Columbia University*

AUTOMATIC DEBIASED MACHINE LEARNING OF CAUSAL AND STRUCTURAL EFFECTS

VICTOR CHERNOZHUKOV
Department of Economics, Massachusetts Institute of Technology

WHITNEY K. NEWEY
Department of Economics, Massachusetts Institute of Technology and NBER

RAHUL SINGH
Department of Economics, Massachusetts Institute of Technology

---

Automatic Debiased Machine Learning for Smooth Functionals of Nonparametric M-Estimands

Lars van der Laan[*1,2], Aurélien Bibaut[2], Nathan Kallus[2,3], and Alex Luedtke[1]

[1]Department of Statistics, University of Washington
[2]Netflix Research
[3]Cornell Tech, Cornell University

— Double Machine Learning (DML):

1. Estimate g_0 by $\hat{g}$.

2. Estimate $\theta_\star$ by $\arg\min_{\theta\in\Theta} \mathbb{E}_{Z\sim\mathbb{P}}[\ell(\theta, \hat{g}; Z)]$.

Question:

How $\hat{g}$ influence the estimation of $\theta_\star$?

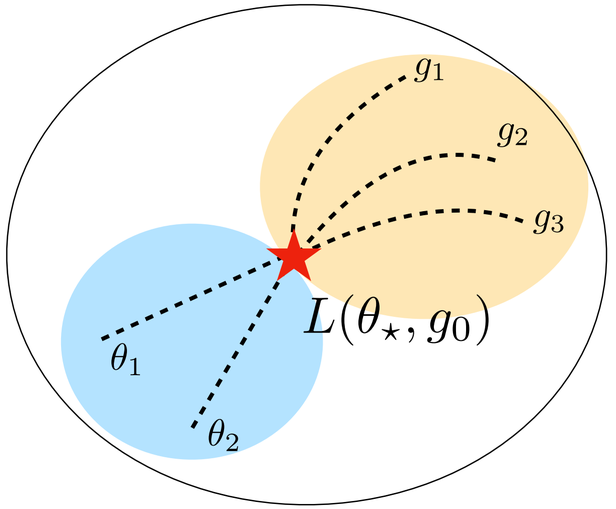
— It depends on the orthogonality of the loss function $\ell(\theta, g; Z)$.

Neyman Orthogonality

Definition 2 (Neyman Orthogonality). *For $\Theta' \subseteq \Theta$, the population loss $L(\theta, g) = \mathbb{E}_{Z\sim\mathbb{P}}[\ell(\theta, g; Z)]$ is Neyman orthogonal at $(\theta_\star, g_0)$ over $\Theta' \times \mathcal{G}'$ if*

$$D_g D_\theta L(\theta_\star, g_0)\big[\theta - \theta_\star,\, g - g_0\big] = 0 \qquad \text{for all } (\theta, g) \in \Theta' \times \mathcal{G}'.$$

- D_θ and D_g are the directional derivative.
- $D_\theta L(\theta_\star, g_0)[\theta - \theta_\star] = \langle \nabla_\theta L(\theta_\star, g_0), \theta - \theta_\star \rangle$.



Orthogonalization — Visualization of the Loss Contour

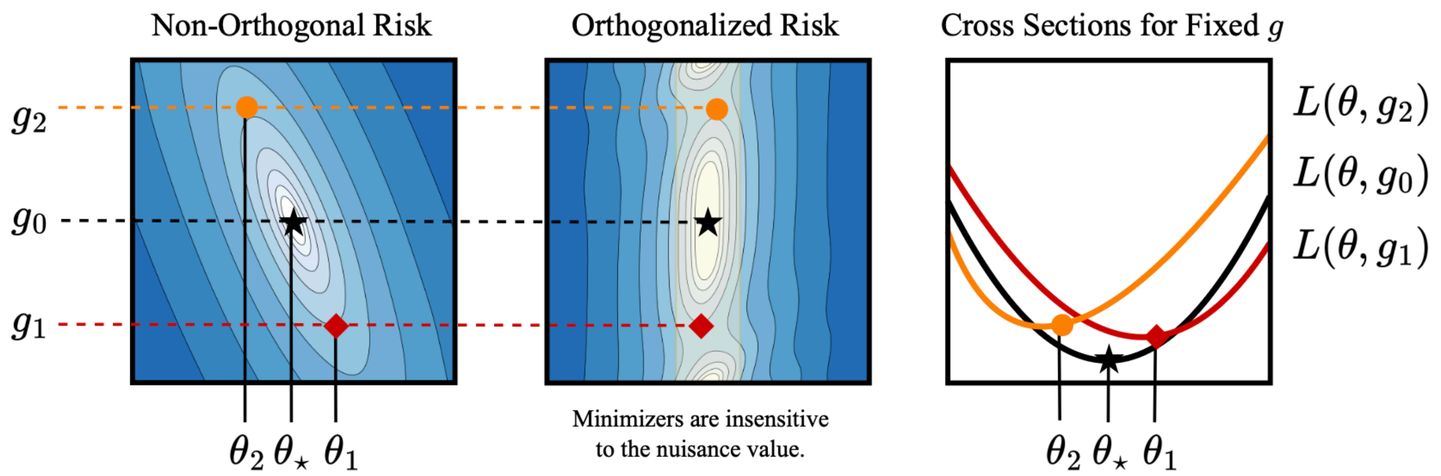


Figure 1: **Illustration of Neyman Orthogonalization.** The first two panels are contour plots of the risk function $L(\theta, g)$, where θ varies on the x-axis and g varies on the y-axis. For the orthogonalized risk (center) the contours are approximately axis-aligned. The right plot shows the cross sections of the non-orthogonal risk when fixing $g = g_0, g_1, g_2$. Due to non-orthogonality, the minimizers θ_1 and θ_2 shown in the first and third plots may drift significantly from $\theta_\star$. In contrast, the minimizers in the center plot are less sensitive to the choice of g.

Example 1: Partially Linear Model

Let $Z = (X, W, Y) \sim \mathbb{P}$ such that

$$Y = \langle \theta_0, X \rangle + g_0(W) + \epsilon, \text{ where } \mathbb{E}[\epsilon \mid X, W] = 0.$$

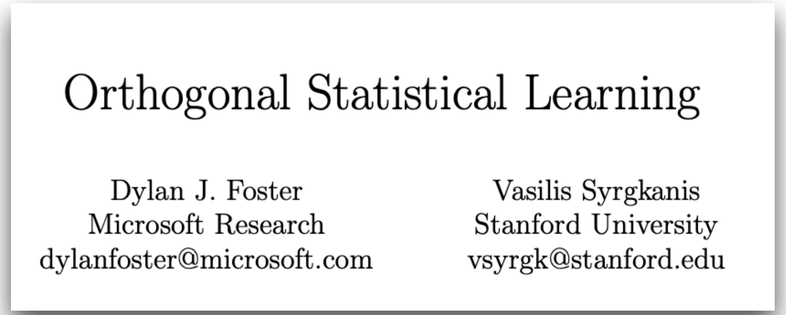
Non-orthogonal loss: $\mathcal{G} = L_2(\mathbb{P}_W)$ $\ell(\theta, g; z) = \frac{1}{2}(y - g(w) - \langle \theta, x \rangle)^2.$

Neyman orthogonal loss: $\ell_\perp(\theta, g; z) = \frac{1}{2}(y - g_Y(w) - \langle \theta, x - g_X(w) \rangle)^2.$

$g_{0,Y}(w) := \mathbb{E}_{\mathbb{P}}[Y \mid W = w]$ and $g_{0,X}(w) := \mathbb{E}_{\mathbb{P}}[X \mid W = w]$,

It is NOT always possible to construct a Neyman orthogonal loss!

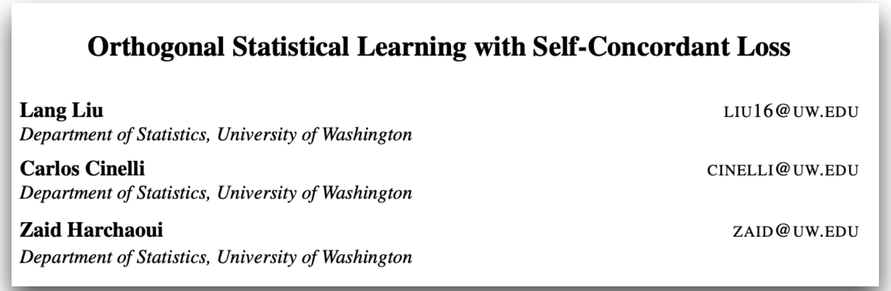
Orthogonal Statistical Learning (OSL)



Orthogonal Statistical Learning

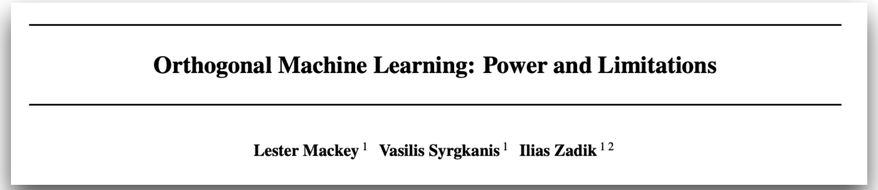
Dylan J. Foster
Microsoft Research
dylanfoster@microsoft.com

Vasilis Syrgkanis
Stanford University
vsyrgk@stanford.edu



Orthogonal Statistical Learning with Self-Concordant Loss

Lang Liu LIU16@UW.EDU
Department of Statistics, University of Washington
Carlos Cinelli CINELLI@UW.EDU
Department of Statistics, University of Washington
Zaid Harchaoui ZAID@UW.EDU
Department of Statistics, University of Washington



Orthogonal Machine Learning: Power and Limitations

Lester Mackey [1] **Vasilis Syrgkanis** [1] **Ilias Zadik** [1 2]

Identify the loss function $\ell(\theta, g; Z)$ as
the Neyman orthogonal loss or the non-orthogonal loss.

Previous work in DML/OSL:

— Empirical Risk Minimization (ERM)

Approximate $\theta_\star$

by $\arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, \hat{g}; Z_i)$.

Our contribution:

— Stochastic Approximation

Approximate $\theta_\star$

by $\theta^{(n)} = \theta^{(n-1)} - \eta S(\theta^{(n-1)}, \hat{g}; Z_n), \theta^{(0)} \in \Theta.$

Algorithm 1.

1. $\hat{g} \in (\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ is a nuisance estimator.
2. $S(\theta, g; z) = \nabla_\theta \ell(\theta, g; z)$ is the gradient.
3. Let $\mathcal{D}_n := (Z_i)_{i=1}^n$ i.i.d. drawn from $\mathbb{P}$.
4. Estimate $\theta_\star$ by SGD updates $\theta^{(n)} = \theta^{(n-1)} - \eta S(\theta^{(n-1)}, \hat{g}; Z_n), \theta^{(0)} \in \Theta$.

Theorem 1. Let $S(\theta, g; z) = \nabla_\theta \ell(\theta, g; z)$. Then for $\hat{g}$ and an appropriate learning rate η, it holds that

$$\mathbb{E}_{\mathcal{D}_n}[\|\theta^{(n)} - \theta_\star\|_2^2] \lesssim (1 - \mu\eta/2)^n + \|\hat{g} - g_0\|_{\mathcal{G}}^2 + \eta.$$

In addition, if L is Neyman orthogonal at $(\theta_\star, g_0)$, then

$$\mathbb{E}_{\mathcal{D}_n}[\|\theta^{(n)} - \theta_\star\|_2^2] \lesssim (1 - \mu\eta/2)^n + \|\hat{g} - g_0\|_{\mathcal{G}}^4 + \eta.$$

Can we orthogonalize the stochastic gradient oracle? Yes!

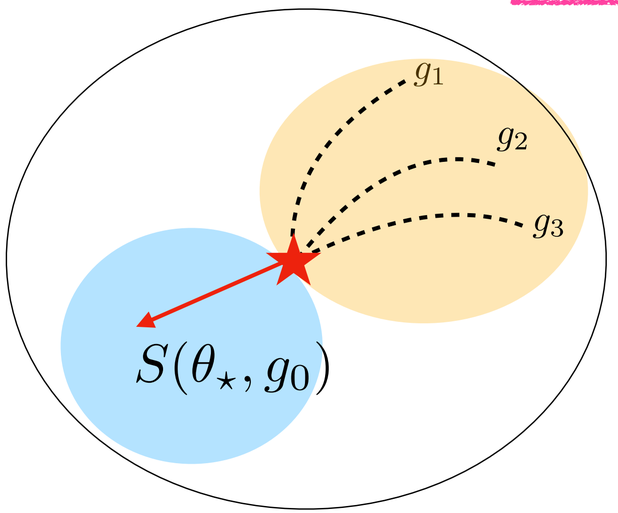
Neyman Orthogonality: the Gradient Level

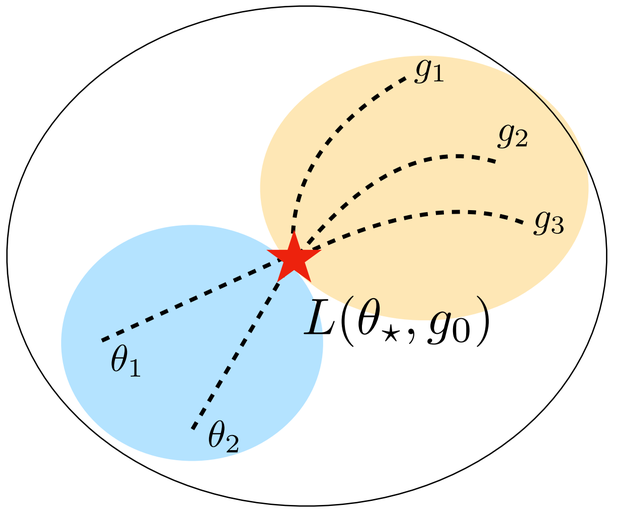
Definition 2 (Neyman Orthogonality). *For $\Theta' \subseteq \Theta$, the population loss $L(\theta, g) = \mathbb{E}_{Z\sim\mathbb{P}}[\ell(\theta, g; Z)]$ is Neyman orthogonal at $(\theta_\star, g_0)$ over $\Theta' \times \mathcal{G}'$ if*

$$D_g D_\theta L(\theta_\star, g_0)\big[\theta - \theta_\star,\ g - g_0\big] = 0 \qquad \text{for all } (\theta, g) \in \Theta' \times \mathcal{G}'.$$

The population gradient oracle $S(\theta, g) = \mathbb{E}_{Z\sim\mathbb{P}}[S(\theta, g; Z)]$ is Neyman orthogonal at $(\theta_\star, g_0)$ over $\mathcal{G}' \subseteq \mathcal{G}$ if

$$D_g S(\theta_\star, g_0)\big[g - g_0\big] = 0 \qquad \text{for all } g \in \mathcal{G}'.$$





Orthogonalization — A Gradient Perspective

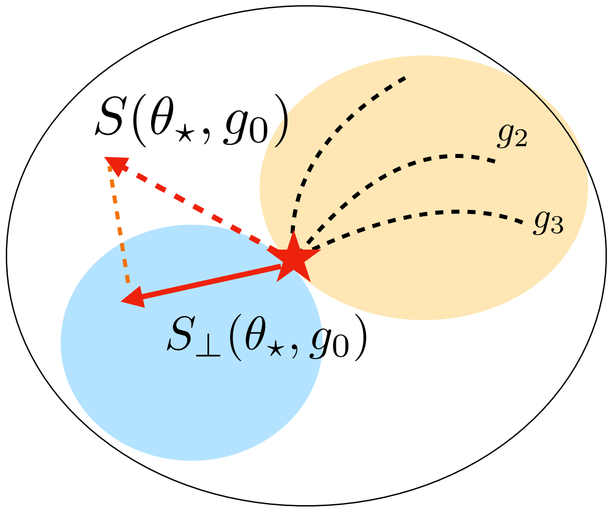
Consider $\ell(\theta, g; z) = -\log p_{\theta,g}(z)$ with $\mathcal{G} = \mathbb{R}^k$.

1. Compute the orthogonalizing operator:

$$\Gamma_0 = \arg\min_{\Gamma \in \mathbb{R}^{d \times k}} \mathbb{E}_{\mathbb{P}} \left[\|\nabla_\theta \ell(\theta_\star, g_0; Z) - \Gamma \nabla_g \ell(\theta_\star, g_0; Z)\|_2^2 \right].$$

2. Remove the projection of the gradient onto the nuisance tangent space:

$$S_\perp(\theta, g; z) = \nabla_\theta \ell(\theta, g; z) - \Gamma_0 \nabla_g \ell(\theta, g; z).$$



This orthogonalization can be Generalize to other losses with possibly infinite-dimensional nuisance space $\mathcal{G}$.

Algorithm 2

1. $\hat{g} \in (\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ is a nuisance estimator.
2. $\hat{\Gamma}$ is the operator estimator for Γ_0.
3. $\hat{S}_\perp(\theta, g; z) = \nabla_\theta \ell(\theta, g; z) - \hat{\Gamma} \nabla_g \ell(\theta, g; z)$.
4. Let $\mathcal{D}_n := (Z_i)_{i=1}^n$ i.i.d. drawn from $\mathbb{P}$.
5. Estimate $\theta_\star$ by SGD updates $\theta^{(n)} = \theta^{(n-1)} - \eta S_\perp(\theta^{(n)}, \hat{g}; Z_n), \theta^{(0)} \in \Theta$.

Theorem 2. Let $S(\theta, g; z) = \hat{S}_\perp(\theta, g; z)$ as the approximated NO gradient by using an estimated operator $\hat{\Gamma}$. Then for $\hat{g}$ and an appropriate learning rate η, it holds that

$$\mathbb{E}_{\mathcal{D}_n}[\|\theta^{(n)} - \theta_\star\|_2^2] \lesssim (1 - \mu\eta/2)^n + \|\hat{g} - g_0\|_{\mathcal{G}}^4 + \|\hat{g} - g_0\|_{\mathcal{G}}^2 \cdot \|\hat{\Gamma} - \Gamma_0\|^2 + \eta.$$

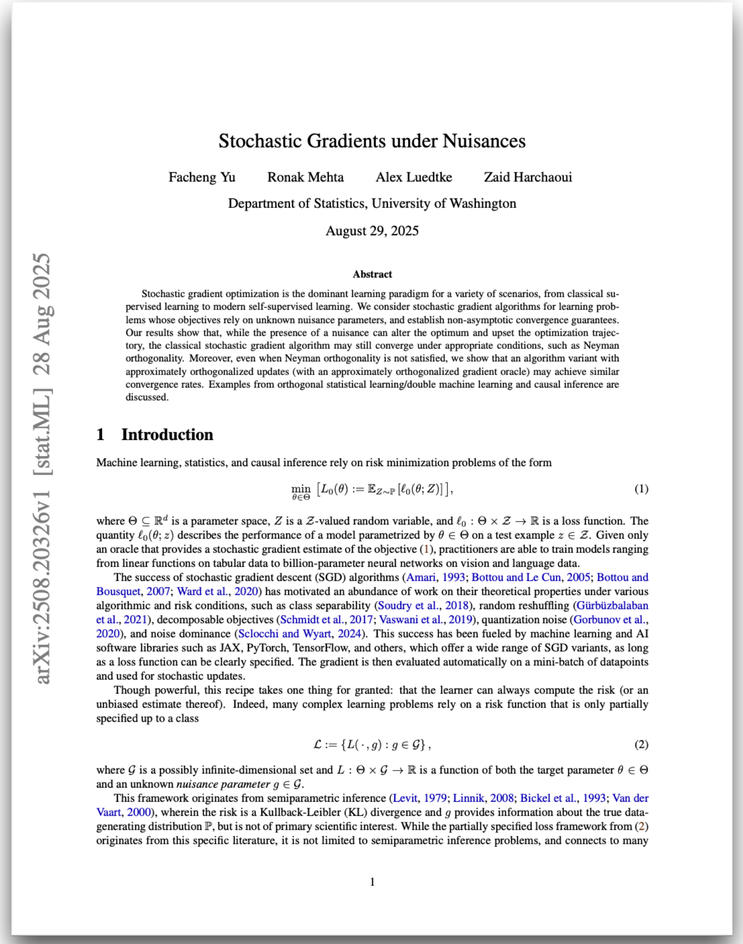
Improve $\|\hat{g} - g_0\|_{\mathcal{G}}^2$ by the cross product $\|\hat{g} - g_0\|_{\mathcal{G}}^2 \cdot \|\hat{\Gamma} - \Gamma_0\|^2$!

Summary

- We studied the learning problem under the nuisance, which applies to, but is not limited to, causal inference, distributionally robust optimization, profile likelihood method.
- We provided non-asymptotic convergence guarantee of stochastic gradient algorithms instead of empirical risk minimization for learning problem under nuisance.
- We designed an approximated orthogonal gradient oracle for non-orthogonal losses to make SGD insensitive to nuisance estimation error.

Extensions

- We discussed how interleaving the nuisance updates and the target updates can improve the estimation performance.
- We discussed several variants of SGD including SGD with momentum, averaged SGD, and Adam.



Stochastic Gradients under Nuisances

Facheng Yu Ronak Mehta Alex Luedtke Zaid Harchaoui

Department of Statistics, University of Washington

August 29, 2025

Abstract

Stochastic gradient optimization is the dominant learning paradigm for a variety of scenarios, from classical supervised learning to modern self-supervised learning. We consider stochastic gradient algorithms for learning problems whose objectives rely on unknown nuisance parameters, and establish non-asymptotic convergence guarantees. Our results show that, while the presence of a nuisance can alter the optimum and upset the optimization trajectory, the classical stochastic gradient algorithm may still converge under appropriate conditions, such as Neyman orthogonality. Moreover, even when Neyman orthogonality is not satisfied, we show that an algorithm variant with approximately orthogonalized updates (with an approximately orthogonalized gradient oracle) may achieve similar convergence rates. Examples from orthogonal statistical learning/double machine learning and causal inference are discussed.

1 Introduction

Machine learning, statistics, and causal inference rely on risk minimization problems of the form

$$\min_{\theta \in \Theta} \left[L_0(\theta) := \mathbb{E}_{Z \sim \mathbb{P}} [\ell_0(\theta; Z)] \right], \tag{1}$$

where $\Theta \subseteq \mathbb{R}^d$ is a parameter space, Z is a $\mathcal{Z}$-valued random variable, and $\ell_0 : \Theta \times \mathcal{Z} \to \mathbb{R}$ is a loss function. The quantity $\ell_0(\theta; z)$ describes the performance of a model parametrized by $\theta \in \Theta$ on a test example $z \in \mathcal{Z}$. Given only an oracle that provides a stochastic gradient estimate of the objective (1), practitioners are able to train models ranging from linear functions on tabular data to billion-parameter neural networks on vision and language data.

The success of stochastic gradient descent (SGD) algorithms (Amari, 1993; Bottou and Le Cun, 2005; Bottou and Bousquet, 2007; Ward et al., 2020) has motivated an abundance of work on their theoretical properties under various algorithmic and risk conditions, such as class separability (Soudry et al., 2018), random reshuffling (Gürbüzbalaban et al., 2021), decomposable objectives (Schmidt et al., 2017; Vaswani et al., 2019), quantization noise (Gorbunov et al., 2020), and noise dominance (Sclocchi and Wyart, 2024). This success has been fueled by machine learning and AI software libraries such as JAX, PyTorch, TensorFlow, and others, which offer a wide range of SGD variants, as long as a loss function can be clearly specified. The gradient is then evaluated automatically on a mini-batch of datapoints and used for stochastic updates.

Though powerful, this recipe takes one thing for granted: that the learner can always compute the risk (or an unbiased estimate thereof). Indeed, many complex learning problems rely on a risk function that is only partially specified up to a class

$$\mathcal{L} := \{L(\cdot, g) : g \in \mathcal{G}\}, \tag{2}$$

where $\mathcal{G}$ is a possibly infinite-dimensional set and $L : \Theta \times \mathcal{G} \to \mathbb{R}$ is a function of both the target parameter $\theta \in \Theta$ and an unknown *nuisance parameter* $g \in \mathcal{G}$.

This framework originates from semiparametric inference (Levit, 1979; Linnik, 2008; Bickel et al., 1993; Van der Vaart, 2000), wherein the risk is a Kullback-Leibler (KL) divergence and g provides information about the true data-generating distribution $\mathbb{P}$, but is not of primary scientific interest. While the partially specified loss framework from (2) originates from this specific literature, it is not limited to semiparametric inference problems, and connects to many

NeurIPS 2025 (to appear)

Thank you!

Paper

Appendix

Interleaving nuisance and target updates

1. Estimate $\hat{g} \in (\mathcal{G}, \|\cdot\|_{\mathcal{G}})$ using $W_1, \ldots, W_m$ i.i.d. drawn from $\mathbb{Q}$.
2. Do $\theta^{(n)} = \theta^{(n-1)} - \eta S(\theta^{(n-1)}, \hat{g}; Z_n), \theta^{(0)} \in \Theta$. $\mathcal{D}_n := (Z_i)_{i=1}^n$ i.i.d. drawn from $\mathbb{P}$.
3. Observe new samples from $\mathbb{Q}$ and $\mathbb{P}$ and repeat 1 and 2.

Proposition 22. *Suppose that $\hat{g}^{(m)}$ satisfies (102) and that $\hat{g}^{(m)} \in \mathcal{G}_r(g_0)$ and $\theta^{(m,t)} \in \Theta$ almost surely for all $m \geq 1$ and $0 \leq t \leq n$. Under Asm. 3, it holds that*

$$\mathbb{E}_{\mathcal{D}_{mn} \cup \mathcal{S}_m \sim \mathbb{P}^{mn} \otimes \mathbb{Q}^m}[\|\theta^{(m,n)} - \theta_\star\|_2^2] \lesssim \left(1 - \frac{\mu\eta}{2}\right)^{mn} \|\theta^{(0)} - \theta_\star\|_2^2 + m \exp\left(-\frac{\mu\eta nm}{4}\right) + (m^{-\frac{2\alpha-1}{2\alpha}} + \eta)((\eta n)^{-1} + 1).$$

$$\mathbb{E}_{\mathcal{S}_m \sim \mathbb{Q}^m}\left[\|\hat{g}^{(m)} - g_0\|_{\mathcal{G}}^2\right] \leq C m^{-\frac{2\alpha-1}{2\alpha}}.$$

In addition, when $(\eta n)^{-1} = \mathcal{O}(1)$, it holds that

$$\mathbb{E}_{\mathcal{D}_{mn} \cup \mathcal{S}_m \sim \mathbb{P}^{mn} \otimes \mathbb{Q}^m}[\|\theta^{(m,n)} - \theta_\star\|_2^2] \lesssim \left(1 - \frac{\mu\eta}{2}\right)^{mn} \|\theta^{(0)} - \theta_\star\|_2^2 + m^{-\frac{2\alpha-1}{2\alpha}} + n^{-1} + \eta.$$

Averaged SGD

Example 5 (Averaged SGD). Let $\beta_n = 1/n$ and $\alpha_n = \eta(1 - \beta_{n+1})$ for all $n \geq 1$. The momentum updates implied by this sequence are

$$m^{(n+1)} = \frac{1}{n} m^{(n)} + S^{(n)} \text{ and } \bar{\theta}^{(n+1)} = \bar{\theta}^{(n)} - \eta \left(1 - \frac{1}{n+1}\right) m^{(n)},$$

which implies that $\bar{\theta}^{(n+1)}$ is the averaged SGD such that

$$\bar{\theta}^{(n+1)} = \frac{1}{n+1} \sum_{t=0}^{n} \theta^{(t)}. \tag{128}$$

Proposition 24 (Convergence rate of averaged SGD)**.** *Consider the partially linear model and the non-orthogonal loss $\ell(\theta, g; z)$ in Appx. B.1.2. Define $\mathcal{D}_n = (Z_1, \ldots, Z_n)$, sampled from the product measure $\mathbb{P}^n$. Choose the gradient oracle $S^{(n)}$ to be the score $S_\theta(\theta, \hat{g}; Z_n)$ where $\hat{g}$ is estimated independently of $\mathcal{D}_n$. Let $\bar{\theta}^{(n)}$ be the averaged SGD defined in (128). Suppose the same assumptions as Lem. 5. If $0 < \eta < \eta_{\max}$, then*

$$\mathbb{E}_{\mathbb{P}} \left[\|\bar{\theta}^{(n)} - \theta_\star\|_2^2 \right] \lesssim \frac{1}{n} + \|\hat{g} - g_0\|_{\mathcal{G}}^2,$$

where $\eta_{\max} = \sup\{\eta > 0 : tr\left(A^\top \mathbb{E}_{\mathbb{P}}[XX^\top]A\right) - \eta \mathbb{E}_{\mathbb{P}}\left[(X^\top A X)^2\right] > 0, \forall A \in \mathcal{S}(\mathbb{R}^d)\}$ and $\mathcal{S}(\mathbb{R}^d)$ is the set of all $d \times d$ symmetric matrices.