Asymptotic normality and optimality in nonsmooth stochastic optimization

Liwei Jiang

Purdue University, Industrial Engineering

Joint work with Damek Davis and Dmitriy Drusvyatskiy

Oct 6, 2025

Background

CLT: For i.i.d. random variables $X_1, X_2, \ldots$ with mean μ and variance σ^2,

$$\sqrt{k}(\bar{X}_k - \mu) \xrightarrow{w} \mathcal{N}(0, \sigma^2).$$

Background

CLT: For i.i.d. random variables $X_1, X_2, \ldots$ with mean μ and variance σ^2,

$$\sqrt{k}(\bar{X}_k - \mu) \xrightarrow{w} \mathcal{N}(0, \sigma^2).$$

Problem: Find $x^\star$ minimizing

$$F(x) = \mathop{\mathbb{E}}_{z \sim \mathcal{P}}[f(x, z)],$$

where $f(\cdot, z)$ are C^2-smooth and strongly convex.

Background

CLT: For i.i.d. random variables $X_1, X_2, \ldots$ with mean μ and variance σ^2,

$$\sqrt{k}(\bar{X}_k - \mu) \xrightarrow{w} \mathcal{N}(0, \sigma^2).$$

Problem: Find $x^\star$ minimizing

$$F(x) = \mathop{\mathbb{E}}_{z \sim \mathcal{P}}[f(x, z)],$$

where $f(\cdot, z)$ are C^2-smooth and strongly convex.

Algorithms:

Background

CLT: For i.i.d. random variables $X_1, X_2, \dots$ with mean μ and variance σ^2,

$$\sqrt{k}(\bar{X}_k - \mu) \xrightarrow{w} \mathcal{N}(0, \sigma^2).$$

Problem: Find $x^\star$ minimizing

$$F(x) = \mathop{\mathbb{E}}_{z \sim \mathcal{P}}[f(x, z)],$$

where $f(\cdot, z)$ are C^2-smooth and strongly convex.

Algorithms:

- Sample average approximation (SAA):

$$x_k = \operatorname*{argmin}_x \frac{1}{k} \sum_{i=1}^{k} f(x, z_i).$$

Background

CLT: For i.i.d. random variables $X_1, X_2, \ldots$ with mean μ and variance σ^2,

$$\sqrt{k}(\bar{X}_k - \mu) \xrightarrow{w} \mathcal{N}(0, \sigma^2).$$

Problem: Find $x^\star$ minimizing

$$F(x) = \mathop{\mathbb{E}}_{z \sim \mathcal{P}}[f(x, z)],$$

where $f(\cdot, z)$ are C^2-smooth and strongly convex.

Algorithms:

- Sample average approximation (SAA):

$$x_k = \operatorname*{argmin}_x \frac{1}{k} \sum_{i=1}^k f(x, z_i).$$

- Stochastic gradient descent (SGD):

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k, z_k)$$

Guarantee for SGD

Theorem (Ruppert '88)(Polyak–Juditsky '92)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$, then under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, \Sigma), \qquad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

[1] (Huber '67)
[2] (Chen et al 20'), (Zhu et al '23), (Roy-Balasubramanian '23)
[3] (Hájek '70), (Le Cam '71), (Duchi-Ruan '18)

Guarantee for SGD

Theorem (Ruppert '88)(Polyak–Juditsky '92)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$, then under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, \Sigma), \qquad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

and $\Sigma = \nabla^2 F(x^\star)^{-1} \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot \nabla^2 F(x^\star)^{-1}$

[1] (Huber '67)
[2] (Chen et al 20'), (Zhu et al '23), (Roy-Balasubramanian '23)
[3] (Hájek '70), (Le Cam '71), (Duchi-Ruan '18)

Guarantee for SGD

Theorem (Ruppert '88)(Polyak–Juditsky '92)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$, then under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, \Sigma), \qquad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

and $\Sigma = \nabla^2 F(x^\star)^{-1} \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot \nabla^2 F(x^\star)^{-1}$

- Similar results for SAA are known.[1]

[1] (Huber '67)
[2] (Chen et al 20'), (Zhu et al '23), (Roy-Balasubramanian '23)
[3] (Hájek '70), (Le Cam '71), (Duchi-Ruan '18)

Guarantee for SGD

Theorem (Ruppert '88)(Polyak–Juditsky '92)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$, then under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, \Sigma), \qquad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

and $\Sigma = \nabla^2 F(x^\star)^{-1} \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot \nabla^2 F(x^\star)^{-1}$

- Similar results for SAA are known.[1]
- Can estimate Σ online and construct confidence intervals for $x^\star$.[2]

[1] (Huber '67)
[2] (Chen et al 20'), (Zhu et al '23), (Roy-Balasubramanian '23)
[3] (Hájek '70), (Le Cam '71), (Duchi-Ruan '18)

Guarantee for SGD

Theorem (Ruppert '88)(Polyak–Juditsky '92)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$, then under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, \Sigma), \qquad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

and $\Sigma = \nabla^2 F(x^\star)^{-1} \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot \nabla^2 F(x^\star)^{-1}$

- Similar results for SAA are known.[1]
- Can estimate Σ online and construct confidence intervals for $x^\star$.[2]
- Moreover, the covariance matrix Σ is "asymptotically optimal".[3]

[1] (Huber '67)
[2] (Chen et al 20'), (Zhu et al '23), (Roy-Balasubramanian '23)
[3] (Hájek '70), (Le Cam '71), (Duchi-Ruan '18)

Generalization to nonsmooth setting?

Constrained optimization:

$$\min_{x} \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f(x, z)] \qquad \text{Subject to: } x \in \mathcal{X},$$

where $f(\cdot, z), F$ are C^2-smooth.

[4] (Dupacová-Wets '88), (Shapiro '89), (King-Rockafellar '93)
[5] (Duchi-Ruan '18)

Generalization to nonsmooth setting?

Constrained optimization:

$$\min_{x} \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f(x, z)] \qquad \text{Subject to: } x \in \mathcal{X},$$

where $f(\cdot, z), F$ are C^2-smooth.

Prior work:

- SAA has asymptotic normality and it is "optimal".[4]

[4] (Dupacová-Wets '88), (Shapiro '89), (King-Rockafellar '93)
[5] (Duchi-Ruan '18)

Generalization to nonsmooth setting?

Constrained optimization:

$$\min_{x} \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f(x,z)] \qquad \text{Subject to: } x \in \mathcal{X},$$

where $f(\cdot, z), F$ are C^2-smooth.

Prior work:

- SAA has asymptotic normality and it is "optimal".[4]
- No known practical online first-order method is "optimal".

[4] (Dupacová-Wets '88), (Shapiro '89), (King-Rockafellar '93)
[5] (Duchi-Ruan '18)

Generalization to nonsmooth setting?

Constrained optimization:

$$\min_{x} \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f(x,z)] \qquad \text{Subject to: } x \in \mathcal{X},$$

where $f(\cdot, z), F$ are C^2-smooth.

Prior work:

- SAA has asymptotic normality and it is "optimal".[4]
- No known practical online first-order method is "optimal".
  - Dual averaging was shown to be suboptimal.[5]

[4] (Dupacová-Wets '88), (Shapiro '89), (King-Rockafellar '93)
[5] (Duchi-Ruan '18)

Generalization to nonsmooth setting?

Constrained optimization:

$$\min_x \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f(x,z)] \qquad \text{Subject to: } x \in \mathcal{X},$$

where $f(\cdot, z), F$ are C^2-smooth.

Prior work:

- SAA has asymptotic normality and it is "optimal".[4]
- No known practical online first-order method is "optimal".
    - Dual averaging was shown to be suboptimal.[5]
- "Projected SGD" conjectured not asymptotically normal/optimal.[5]

[4] (Dupacová-Wets '88), (Shapiro '89), (King-Rockafellar '93)
[5] (Duchi-Ruan '18)

Generalization to nonsmooth setting?

Constrained optimization:

$$\min_{x} \quad F(x) = \mathop{\mathbb{E}}_{z\in\mathcal{P}}[f(x,z)] \qquad \text{Subject to: } x \in \mathcal{X},$$

where $f(\cdot, z), F$ are C^2-smooth.

Prior work:

- SAA has asymptotic normality and it is "optimal".[4]
- No known practical online first-order method is "optimal".
  - Dual averaging was shown to be suboptimal.[5]
- "Projected SGD" conjectured not asymptotically normal/optimal.[5]

Question:

Is there a gap between offline and first-order online algorithms for constrained optimization?

[4] (Dupacová-Wets '88), (Shapiro '89), (King-Rockafellar '93)
[5] (Duchi-Ruan '18)

Example

Example: Consider solving

$$\min_{x \in \mathbb{R}^3} \mathop{\mathbb{E}}_{z \sim N(-e_3, I)} \langle z, x \rangle = -x_3$$

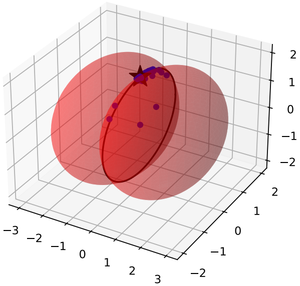
$$\text{subject to: } x \in B_2(e_1) \cap B_2(-e_1)$$

Example

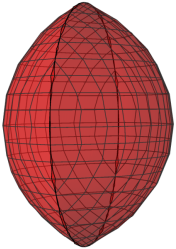
Example: Consider solving

$$\min_{x \in \mathbb{R}^3} \mathop{\mathbb{E}}_{z \sim N(-e_3, I)} \langle z, x \rangle = -x_3$$

$$\text{subject to: } x \in B_2(e_1) \cap B_2(-e_1)$$



(a) Iterates

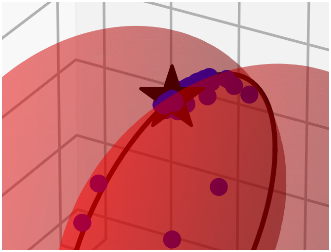


(b) Constraint set

Example

Stochastic projected gradient descent:

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k, z_k)).$$

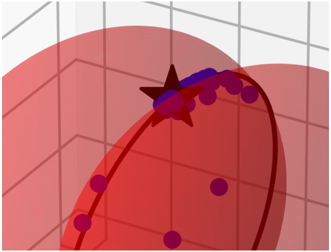


(a) Iterates

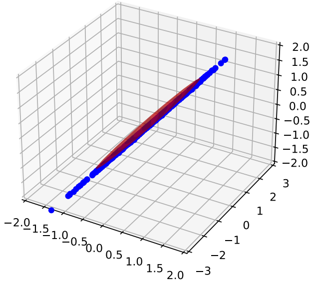
Example

Stochastic projected gradient descent:

$$x_{k+1} = \text{Proj}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k, z_k)).$$



(a) Iterates



(b) $\sqrt{k}(\bar{x}_k - x^\star)$

Example

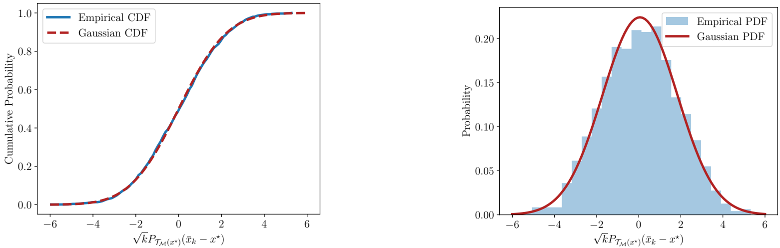
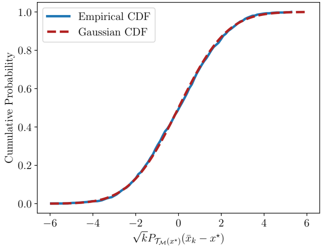


Figure: Empirical vs Gaussian

Example



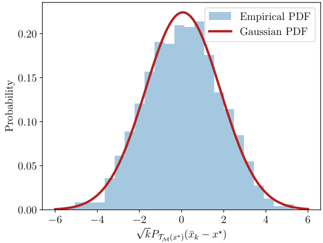


Figure: Empirical vs Gaussian

Observations:

- $\sqrt{k}(\bar{x}_k - x^\star)$ converges in distribution to a Gaussian.

Example

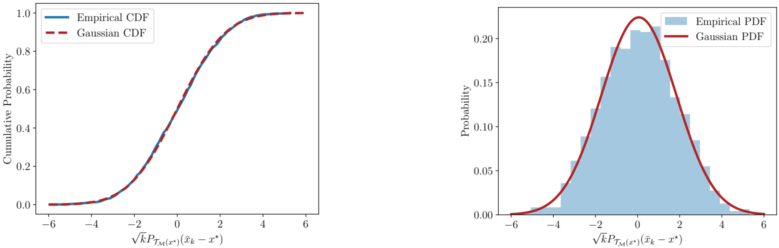
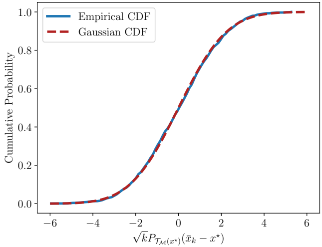


Figure: Empirical vs Gaussian

Observations:

- $\sqrt{k}(\bar{x}_k - x^\star)$ converges in distribution to a Gaussian.
- The covariance matrix is singular.

Example



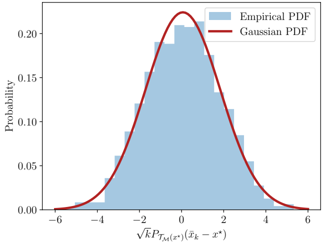


Figure: Empirical vs Gaussian

Observations:

- $\sqrt{k}(\bar{x}_k - x^\star)$ converges in distribution to a Gaussian.
- The covariance matrix is singular.
- The range of the Gaussian is tangent to the circle.

Setting: nonlinear programming

Problem:

$$\min_{x} \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}}[f(x,z)] \qquad \text{Subject to: } g_i(x) \leq 0,$$

where $\{g_i\}_{i\in[m]}$ are smooth. $x^\star$ is the solution.

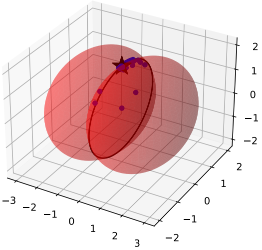
Setting: nonlinear programming

Problem:

$$\min_{x} \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}}[f(x,z)] \qquad \text{Subject to: } g_i(x) \le 0,$$

where $\{g_i\}_{i \in [m]}$ are smooth. $x^\star$ is the solution. Define

$$\mathcal{I} = \{i \colon g_i(x^\star) = 0\}$$
$$\mathcal{M} = \{x \colon g_i(x) = 0, \forall i \in \mathcal{I}\}$$



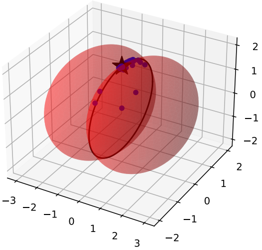
Setting: nonlinear programming

Problem:

$$\min_{x} \; F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f(x,z)] \qquad \text{Subject to: } g_i(x) \leq 0,$$

where $\{g_i\}_{i \in [m]}$ are smooth. $x^\star$ is the solution. Define

$$\mathcal{I} = \{i \colon g_i(x^\star) = 0\}$$
$$\mathcal{M} = \{x \colon g_i(x) = 0, \forall i \in \mathcal{I}\}$$



Standard assumptions:

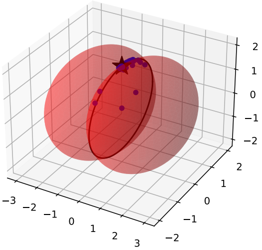
Setting: nonlinear programming

Problem:

$$\min_{x} \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f(x,z)] \qquad \text{Subject to: } g_i(x) \leq 0,$$

where $\{g_i\}_{i \in [m]}$ are smooth. $x^\star$ is the solution. Define

$$\mathcal{I} = \{i \colon g_i(x^\star) = 0\}$$
$$\mathcal{M} = \{x \colon g_i(x) = 0, \forall i \in \mathcal{I}\}$$



Standard assumptions:

- $\{\nabla g_i(x^\star)\}_{i \in \mathcal{I}}$ are linearly independent (LICQ)

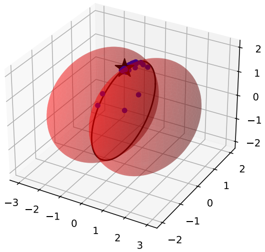
Setting: nonlinear programming

Problem:

$$\min_x \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}}[f(x,z)] \qquad \text{Subject to: } g_i(x) \le 0,$$

where $\{g_i\}_{i\in[m]}$ are smooth. $x^\star$ is the solution. Define

$$\mathcal{I} = \{i\colon g_i(x^\star) = 0\}$$
$$\mathcal{M} = \{x\colon g_i(x) = 0, \forall i \in \mathcal{I}\}$$



Standard assumptions:

- $\{\nabla g_i(x^\star)\}_{i\in\mathcal{I}}$ are linearly independent (LICQ)
- $\lambda_i^\star > 0$ for $i \in \mathcal{I}$ (strict complementarity)

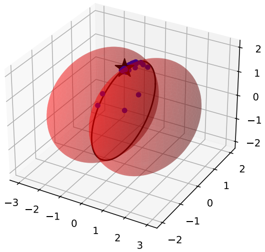
Setting: nonlinear programming

Problem:

$$\min_x \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}}[f(x,z)] \qquad \text{Subject to: } g_i(x) \le 0,$$

where $\{g_i\}_{i\in[m]}$ are smooth. $x^\star$ is the solution. Define

$$\mathcal{I} = \{i\colon g_i(x^\star) = 0\}$$
$$\mathcal{M} = \{x\colon g_i(x) = 0, \forall i \in \mathcal{I}\}$$



Standard assumptions:

- $\{\nabla g_i(x^\star)\}_{i\in\mathcal{I}}$ are linearly independent (LICQ)
- $\lambda_i^\star > 0$ for $i \in \mathcal{I}$ (strict complementarity)
- $u^\top \nabla^2_{xx}\mathcal{L}(x^\star, \lambda^\star)u > 0$ for all nonzero $u \in T_{\mathcal{M}}(x^\star)$. (SSOC)

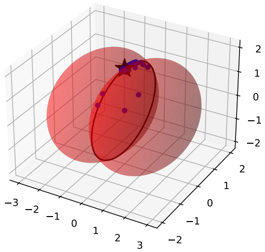
Setting: nonlinear programming

Problem:

$$\min_x \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}}[f(x,z)] \qquad \text{Subject to: } g_i(x) \le 0,$$

where $\{g_i\}_{i\in[m]}$ are smooth. $x^\star$ is the solution. Define

$$\mathcal{I} = \{i : g_i(x^\star) = 0\}$$
$$\mathcal{M} = \{x : g_i(x) = 0, \forall i \in \mathcal{I}\}$$



Standard assumptions:

- $\{\nabla g_i(x^\star)\}_{i\in\mathcal{I}}$ are linearly independent (LICQ)
- $\lambda_i^\star > 0$ for $i \in \mathcal{I}$ (strict complementarity)
- $u^\top \nabla^2_{xx}\mathcal{L}(x^\star, \lambda^\star)u > 0$ for all nonzero $u \in T_{\mathcal{M}}(x^\star)$. (SSOC)

Consequences:

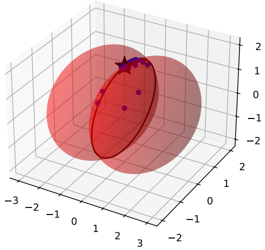
Setting: nonlinear programming

Problem:

$$\min_x \; F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}}[f(x,z)] \qquad \text{Subject to: } g_i(x) \le 0,$$

where $\{g_i\}_{i\in[m]}$ are smooth. $x^\star$ is the solution. Define

$$\mathcal{I} = \{i\colon g_i(x^\star) = 0\}$$
$$\mathcal{M} = \{x\colon g_i(x) = 0, \forall i \in \mathcal{I}\}$$



Standard assumptions:

- $\{\nabla g_i(x^\star)\}_{i\in\mathcal{I}}$ are linearly independent (LICQ)
- $\lambda_i^\star > 0$ for $i \in \mathcal{I}$ (strict complementarity)
- $u^\top \nabla^2_{xx}\mathcal{L}(x^\star, \lambda^\star)u > 0$ for all nonzero $u \in T_{\mathcal{M}}(x^\star)$. (SSOC)

Consequences:

- Locally near $x^\star$, $\mathcal{M}$ is a smooth manifold.

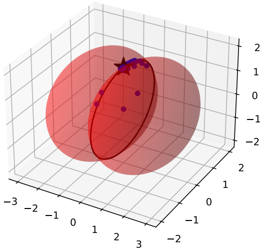
Setting: nonlinear programming

Problem:

$$\min_x \quad F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}}[f(x,z)] \qquad \text{Subject to: } g_i(x) \le 0,$$

where $\{g_i\}_{i\in[m]}$ are smooth. $x^\star$ is the solution. Define

$$\mathcal{I} = \{i \colon g_i(x^\star) = 0\}$$
$$\mathcal{M} = \{x \colon g_i(x) = 0, \forall i \in \mathcal{I}\}$$



Standard assumptions:

- $\{\nabla g_i(x^\star)\}_{i\in\mathcal{I}}$ are linearly independent (LICQ)
- $\lambda_i^\star > 0$ for $i \in \mathcal{I}$ (strict complementarity)
- $u^\top \nabla^2_{xx}\mathcal{L}(x^\star, \lambda^\star)u > 0$ for all nonzero $u \in T_{\mathcal{M}}(x^\star)$. (SSOC)

Consequences:

- Locally near $x^\star$, $\mathcal{M}$ is a smooth manifold.
- for $x \in \mathcal{X}$ near $x^\star$, $F(x) - F(P_{\mathcal{M}}(x)) \gtrsim \mathrm{dist}(x, \mathcal{M})$ (linear growth)

Main idea of our approach

Projected SGD:

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k, z_k)).$$

Main idea of our approach

Projected SGD:

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k, z_k)).$$

Challenge:

$\mathrm{Proj}_{\mathcal{X}}$ is nondifferentiable and nonlinear.

Main idea of our approach

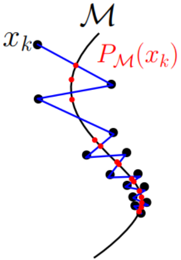
Projected SGD:

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k, z_k)).$$

Challenge:
$\mathrm{Proj}_{\mathcal{X}}$ is nondifferentiable and nonlinear.

Our approach:
Instead of tracking $\{x_k\}$, we consider the

$$\text{shadow sequence}: y_k = P_{\mathcal{M}}(x_k).$$



Main idea of our approach

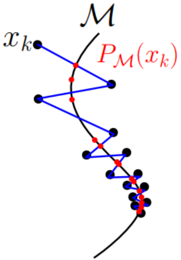
Projected SGD:

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k, z_k)).$$

Challenge:
$\mathrm{Proj}_{\mathcal{X}}$ is nondifferentiable and nonlinear.

Our approach:
Instead of tracking $\{x_k\}$, we consider the

shadow sequence: $y_k = P_{\mathcal{M}}(x_k)$.



Key steps:

Main idea of our approach

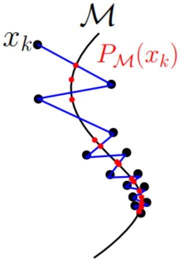
Projected SGD:

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k, z_k)).$$

Challenge:
$\mathrm{Proj}_{\mathcal{X}}$ is nondifferentiable and nonlinear.

Our approach:
Instead of tracking $\{x_k\}$, we consider the

shadow sequence: $y_k = P_{\mathcal{M}}(x_k)$.



Key steps:

- Sharp growth implies x_k reaches $\mathcal{M}$ quickly.

Main idea of our approach

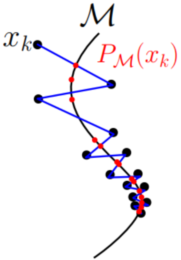
Projected SGD:

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k, z_k)).$$

Challenge:
$\mathrm{Proj}_{\mathcal{X}}$ is nondifferentiable and nonlinear.

Our approach:
Instead of tracking $\{x_k\}$, we consider the

shadow sequence: $y_k = P_{\mathcal{M}}(x_k)$.



Key steps:

- Sharp growth implies x_k reaches $\mathcal{M}$ quickly.
  - $\implies \sqrt{k}(\bar{x}_k - x^\star)$ and $\sqrt{k}(\bar{y}_k - x^\star)$ have same asymp. dist.

Main idea of our approach

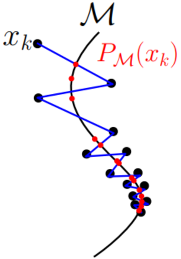
Projected SGD:

$$x_{k+1} = \text{Proj}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k, z_k)).$$

Challenge:
$\text{Proj}_{\mathcal{X}}$ is nondifferentiable and nonlinear.

Our approach:
Instead of tracking $\{x_k\}$, we consider the

shadow sequence: $y_k = P_{\mathcal{M}}(x_k)$.



Key steps:

- Sharp growth implies x_k reaches $\mathcal{M}$ quickly.
  - $\implies \sqrt{k}(\bar{x}_k - x^\star)$ and $\sqrt{k}(\bar{y}_k - x^\star)$ have same asymp. dist.
- The shadow sequence follows the dynamics:

$$y_{k+1} = y_k - \alpha_k \underbrace{\nabla_{\mathcal{M}} f(y_k, z_k)}_{\text{smooth dynamics}} + \underbrace{O(\alpha_k^2)}_{\text{error}}.$$

Main idea of our approach

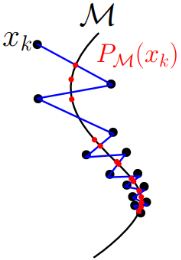
Projected SGD:

$$x_{k+1} = \mathrm{Proj}_{\mathcal{X}}(x_k - \alpha_k \nabla f(x_k, z_k)).$$

Challenge:
$\mathrm{Proj}_{\mathcal{X}}$ is nondifferentiable and nonlinear.

Our approach:
Instead of tracking $\{x_k\}$, we consider the

shadow sequence: $y_k = P_{\mathcal{M}}(x_k)$.



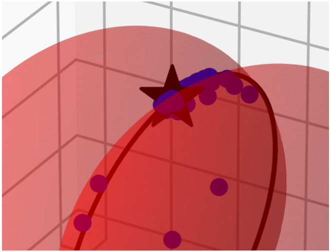
Key steps:

- Sharp growth implies x_k reaches $\mathcal{M}$ quickly.
  - $\implies \sqrt{k}(\bar{x}_k - x^\star)$ and $\sqrt{k}(\bar{y}_k - x^\star)$ have same asymp. dist.
- The shadow sequence follows the dynamics:

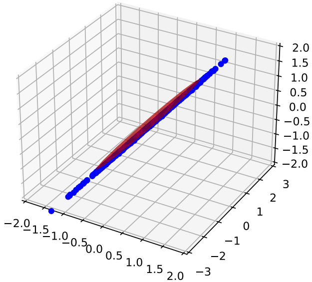
$$y_{k+1} = y_k - \alpha_k \underbrace{\nabla_{\mathcal{M}} f(y_k, z_k)}_{\text{smooth dynamics}} + \underbrace{O(\alpha_k^2)}_{\text{error}}.$$

"Approximate Riemannian SGD"

Illustration



(a) Iterates



(b) $\sqrt{k}(\bar{x}_k - x^\star)$

Main theorem

Theorem (Davis–Drusvyatskiy-J '23)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ and $x_k \to x^\star$, under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, H^\dagger \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot H^\dagger), \quad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

[6] (Duchi-Ruan '18)

Main theorem

Theorem (Davis–Drusvyatskiy-J '23)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ and $x_k \to x^\star$, under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, H^\dagger \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot H^\dagger), \quad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

and $H = P_{T_{\mathcal{M}}(x^\star)} \nabla^2_{xx} \mathcal{L}(x^\star, y^\star) P_{T_{\mathcal{M}}(x^\star)}$

[6] (Duchi-Ruan '18)

Main theorem

Theorem (Davis–Drusvyatskiy-J '23)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ and $x_k \to x^\star$, under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, H^\dagger \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot H^\dagger), \quad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

and $H = P_{T_{\mathcal{M}}(x^\star)} \nabla^2_{xx} \mathcal{L}(x^\star, y^\star) P_{T_{\mathcal{M}}(x^\star)}$

- H is the "Riemannian Hessian".

[6] (Duchi-Ruan '18)

Main theorem

Theorem (Davis–Drusvyatskiy-J '23)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ and $x_k \to x^\star$, under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, H^\dagger \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot H^\dagger), \quad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

and $H = P_{T_{\mathcal{M}}(x^\star)} \nabla^2_{xx} \mathcal{L}(x^\star, y^\star) P_{T_{\mathcal{M}}(x^\star)}$

- H is the "Riemannian Hessian".
- Covariance known to be "optimal".[6]

[6] (Duchi-Ruan '18)

Main theorem

Theorem (Davis–Drusvyatskiy-J '23)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ and $x_k \to x^\star$, under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, H^\dagger \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot H^\dagger), \quad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

and $H = P_{T_{\mathcal{M}}(x^\star)} \nabla^2_{xx} \mathcal{L}(x^\star, y^\star) P_{T_{\mathcal{M}}(x^\star)}$

- H is the "Riemannian Hessian".
- Covariance known to be "optimal".[6]
- Same result holds for Riemannian SGD.

[6] (Duchi-Ruan '18)

Main theorem

Theorem (Davis–Drusvyatskiy-J '23)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ and $x_k \to x^\star$, under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, H^\dagger \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot H^\dagger), \quad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

and $H = P_{T_{\mathcal{M}}(x^\star)} \nabla^2_{xx} \mathcal{L}(x^\star, y^\star) P_{T_{\mathcal{M}}(x^\star)}$

- H is the "Riemannian Hessian".
- Covariance known to be "optimal".[6]
- Same result holds for Riemannian SGD.
  - **Surprising:** Unlike Riemannian SGD, we do not know $\mathcal{M}$.

[6] (Duchi-Ruan '18)

Main theorem

Theorem (Davis–Drusvyatskiy-J '23)

If $\alpha_k = \alpha_0 k^{-\beta}$ for $\beta \in (\frac{1}{2}, 1)$ and $x_k \to x^\star$, under standard noise conditions,

$$\sqrt{k}(\bar{x}_k - x^\star) \xrightarrow{w} \mathcal{N}(0, H^\dagger \cdot \mathrm{Cov}(\nabla f(x^\star, z)) \cdot H^\dagger), \quad \text{where } \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

and $H = P_{T_{\mathcal{M}}(x^\star)} \nabla^2_{xx} \mathcal{L}(x^\star, y^\star) P_{T_{\mathcal{M}}(x^\star)}$

- H is the "Riemannian Hessian".
- Covariance known to be "optimal".[6]
- Same result holds for Riemannian SGD.
  - **Surprising:** Unlike Riemannian SGD, we do not know $\mathcal{M}$.
- Results extend to the stochastic subgradient method and stochastic proximal gradient method

[6] (Duchi-Ruan '18)

Conclusion

[7] (Davis-Drusvyatskiy-J '22)

Conclusion

- Closed the gap between offline and first-order online algorithms for stochastic nonlinear programming.

[7] (Davis-Drusvyatskiy-J '22)

Conclusion

- Closed the gap between offline and first-order online algorithms for stochastic nonlinear programming.
  - Results adapt to nonsmooth stochastic approximation.

[7] (Davis-Drusvyatskiy-J '22)

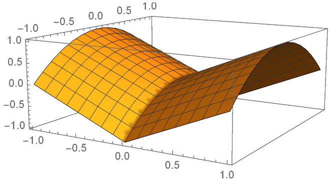
Conclusion

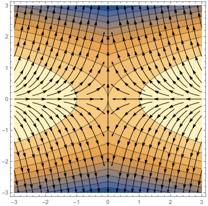
- Closed the gap between offline and first-order online algorithms for stochastic nonlinear programming.
  - Results adapt to nonsmooth stochastic approximation.
- Key idea: shadow sequence $\equiv$ approximate Riemmanian gradient sequence.

[7] (Davis-Drusvyatskiy-J '22)

Conclusion

- Closed the gap between offline and first-order online algorithms for stochastic nonlinear programming.
    - Results adapt to nonsmooth stochastic approximation.
- Key idea: shadow sequence $\equiv$ approximate Riemmanian gradient sequence.
    - Our related work used shadow sequence shows that SGD escapes saddle points of nonsmooth/constrained problems[7]





[7] (Davis-Drusvyatskiy-J '22)

More examples

Unconstrained examples:

- The objective itself can be nonsmooth:

$$\min_x F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f(x,z)] + \lambda \|x\|_1.$$

- Generic semi-algebraic functions

More examples

Unconstrained examples:

- The objective itself can be nonsmooth:

$$\min_x F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f(x,z)] + \lambda \|x\|_1.$$

- Generic semi-algebraic functions

Stochastic variational inequalities:
We consider the task of finding a solution $x^\star$ of the inclusion

$$0 \in \mathop{\mathbb{E}}_{z \in \mathcal{P}} [A(x,z)] + N_{\mathcal{X}}(x),$$

where $A(\cdot, z)$ is a smooth map for almost every $z \sim \mathcal{P}$.

More examples

Unconstrained examples:

- The objective itself can be nonsmooth:

$$\min_x F(x) = \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f(x,z)] + \lambda \|x\|_1.$$

- Generic semi-algebraic functions

Stochastic variational inequalities:
We consider the task of finding a solution $x^\star$ of the inclusion

$$0 \in \mathop{\mathbb{E}}_{z \in \mathcal{P}} [A(x,z)] + N_{\mathcal{X}}(x),$$

where $A(\cdot, z)$ is a smooth map for almost every $z \sim \mathcal{P}$.

Stochastic equilibrium problem:
Nash equilibria $x^\star = (x_1^\star, \ldots, x_m^\star)$ of stochastic games are solutions of the system

$$x_j^\star \in \operatorname*{argmin}_{x_j \in \mathcal{X}_j} \mathop{\mathbb{E}}_{z \in \mathcal{P}} [f_j(x,z)], \qquad \text{for all } j = 1, \ldots, m.$$

If we let $A(x,z)$ be a map that $[A(x,z)]_j = \nabla_{x_j} f_j(x,z)$, and $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_m$, the problem becomes stochastic variational inequalities.