Optimal sampling for SGD

Journée scientifique SMAI-SIGMA — December 2023

Philipp Trunschke, Anthony Nouy, Robert Gruhlke

Funding: COFNET (ANR-DFG)

Nantes Université

CENTRALE NANTES

VERITAS IUSTITIA LIBERTAS
FREIE UNIVERSITÄT BERLIN

Berlin Mathematics Research Center
MATH+

Laboratoire de Mathématiques Jean Leray

Structure

- Motivation
- Setting
- Assumptions
- Results
- Experiments

Motivation

Setting

- Let ρ be a probability measure on $\mathscr{X}$.
- Let $\mathscr{M}$ be a model class of functions on $\mathscr{X}$.
- Consider the problem

$$u^\star = \arg\min_{v \in \mathscr{M}} \mathscr{L}(v) \quad \text{with} \quad \mathscr{L}(v) := \int \ell(v; x)\, \mathrm{d}\rho(x)\,.$$

Generalisation

- If $\mathscr{L}$ is replaced by a MC estimate $\mathscr{L}_n$ with sample size n,

$$u_n^\star := \arg\min_{v \in \mathscr{M}} \mathscr{L}_n(v)$$

- This ensues a generalisation error.
- Bounding this error requires strong assumptions.

Generalisation error bounds

- Suppose that $\mathscr{M}$ is compact.
- Suppose ℓ is bounded and $\ell(\,\bullet\,; x)$ is Lipschitz on $\mathscr{M}$ for all $x \in \mathscr{X}$.
- Then

$$\mathscr{L}(u_n^\star) \leq \mathscr{L}(u^\star) + \mathcal{O}(n^{-1/2})\,.$$

Generalisation error bounds

- Suppose that $\mathscr{M}$ is compact.
- Suppose ℓ is bounded and $\ell(\bullet\,; x)$ is Lipschitz on $\mathscr{M}$ for all $x \in \mathscr{X}$.
- Then

$$\mathscr{L}(u_n^\star) \leq \mathscr{L}(u^\star) + \boxed{\mathcal{O}(n^{-1/2})}.$$ slow convergence

We want to use a minimal amount of samples!

Generalisation error bounds

strong requirements

- Suppose that $\mathscr{M}$ is compact.
- Suppose ℓ is bounded and $\ell(\,\bullet\,; x)$ is Lipschitz on $\mathscr{M}$ for all $x \in \mathscr{X}$.
- Then

$$\mathscr{L}(u_n^\star) \leq \mathscr{L}(u^\star) + \boxed{\mathscr{O}(n^{-1/2})}.$$

slow convergence

We want to use a minimal amount of samples!

Compactness may require regularisation → changes the minimum

Idea

- Optimise the true loss $\mathscr{L}$ on the manifold of functions $\mathscr{M}$.
  - No generalisation error!

Thank you for your attention !

Idea

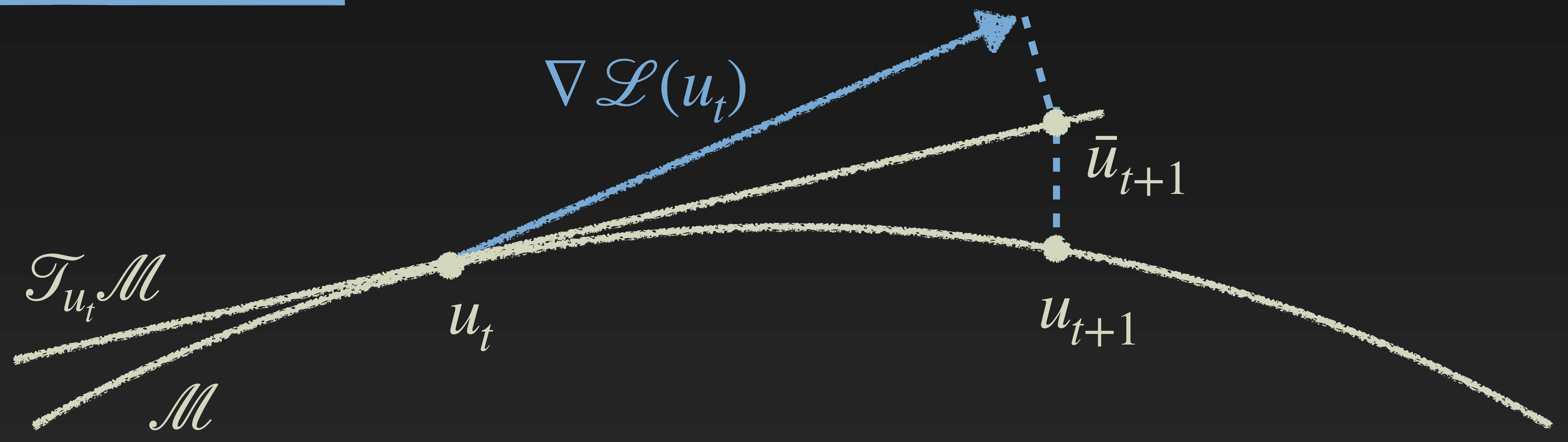
How?

- Optimise the true loss $\mathscr{L}$ on the manifold of functions $\mathscr{M}$.

Idea

How?

- Optimise the true loss $\mathscr{L}$ on the manifold of functions $\mathscr{M}$.

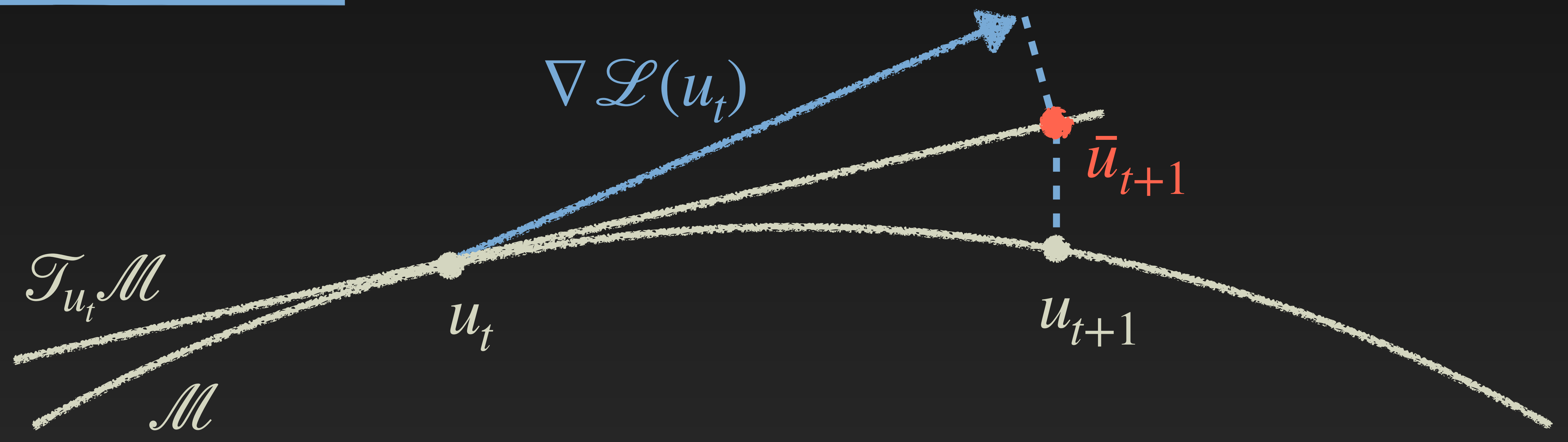


- Perform a Riemannian type optimisation.
- Replace the optimal projection of $\nabla\mathscr{L}(v)$ onto the tangent space with a quasi-optimal least squares projection.

Idea

How?

- Optimise the true loss $\mathscr{L}$ on the manifold of functions $\mathscr{M}$.



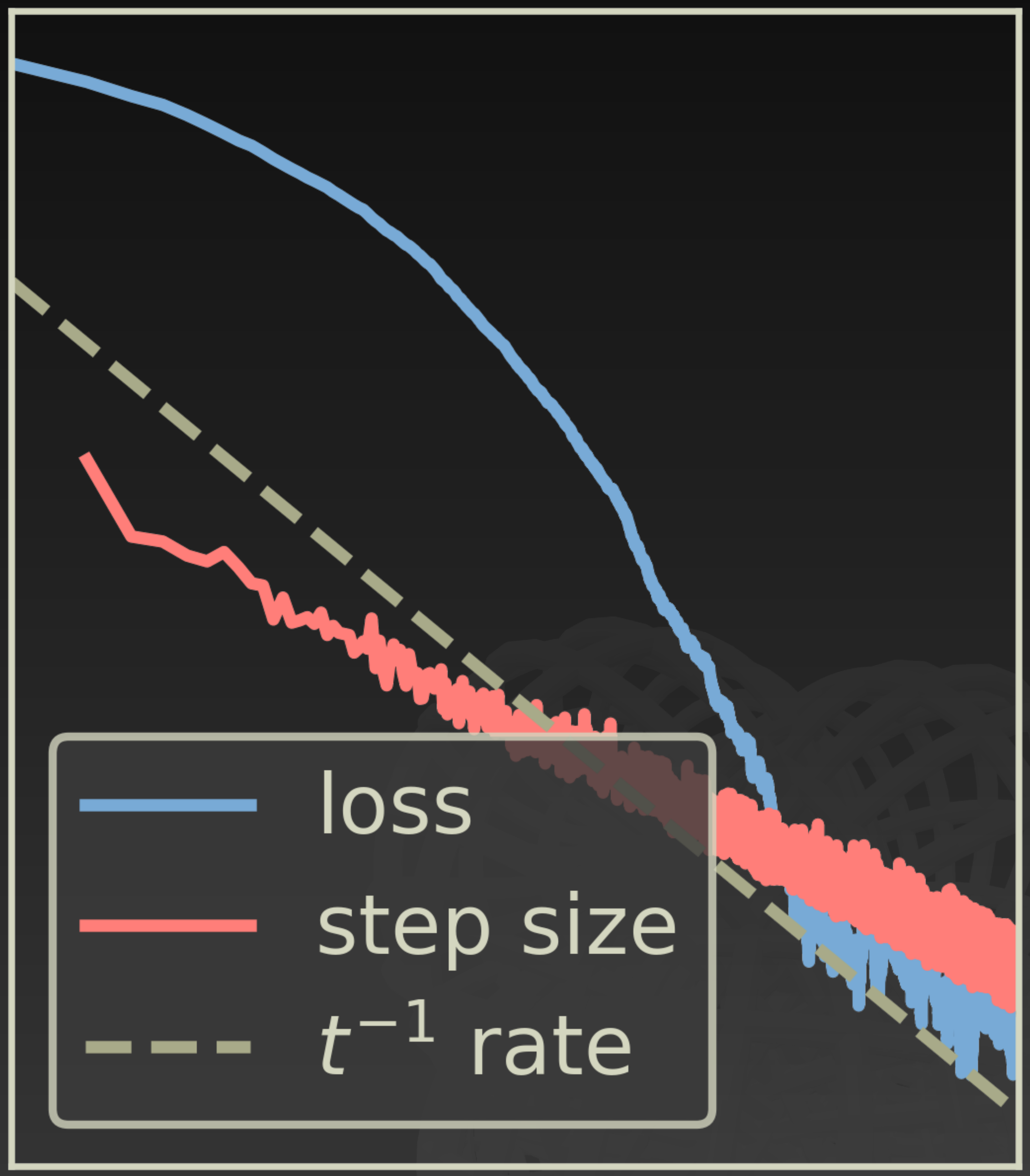
- Perform a Riemannian type optimisation.
- Replace the optimal projection of $\nabla \mathscr{L}(v)$ onto the tangent space with a quasi-optimal least squares projection.

Results

- Almost sure convergence

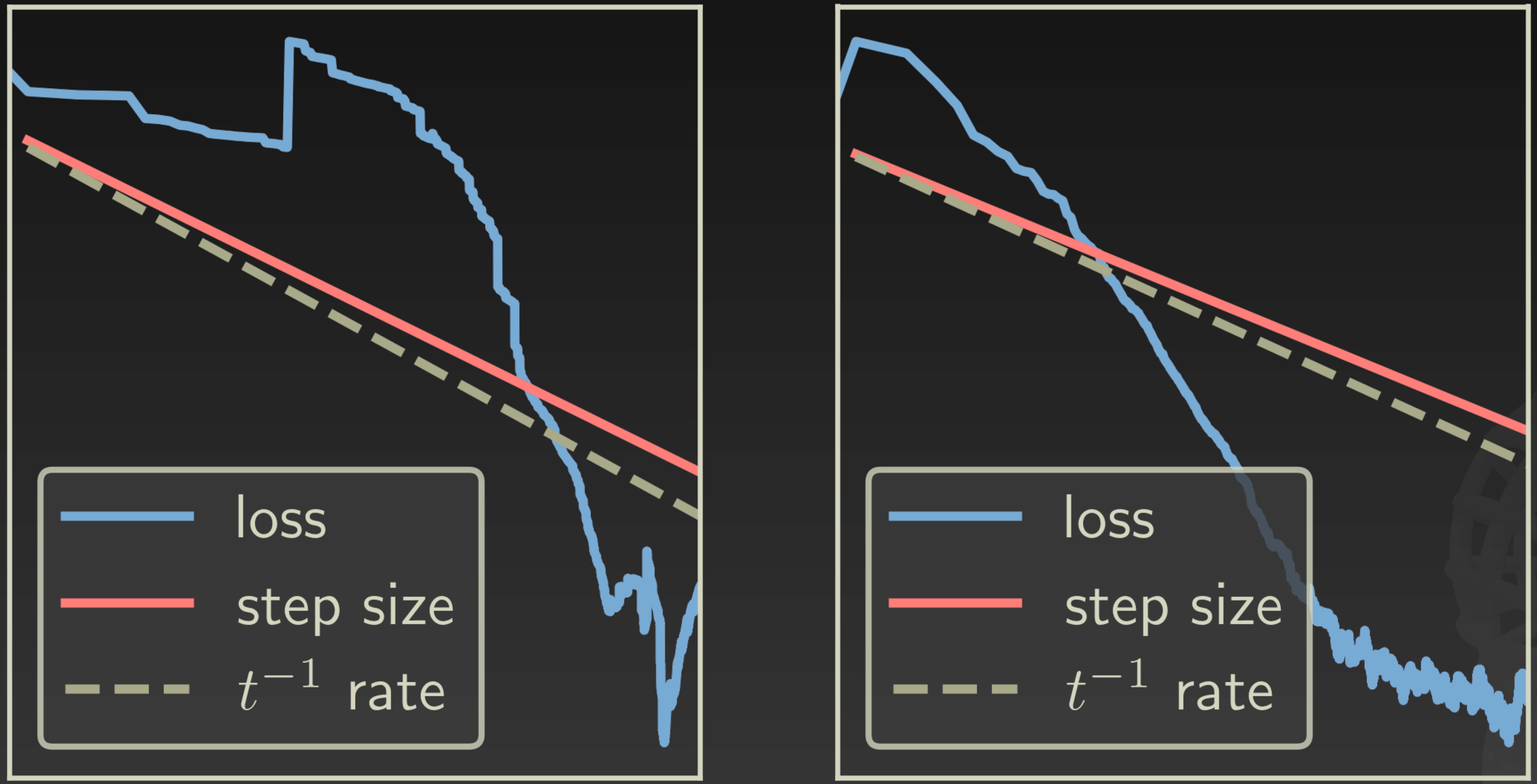
Results

- Almost sure convergence
- Two regimes of convergence:
  - exponential GD rates with large gradients
  - classical SGD rates afterwards



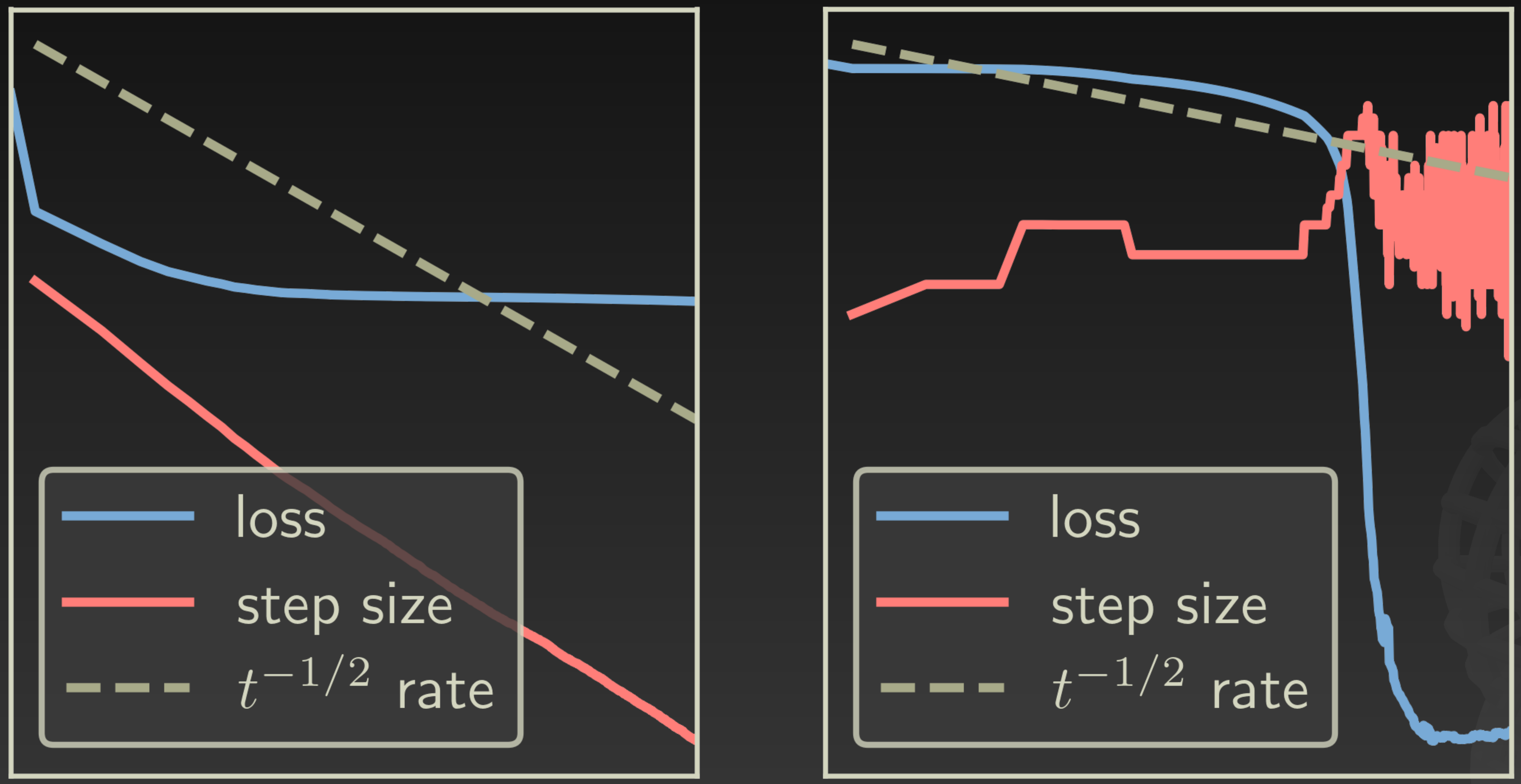
Results

- With optimal sampling, we can control the variance.



Results

- We can converge to the local minimum.



Interlude

Why is this a good idea?

- General regression needs many samples.
- But linear least squares does not!
- Consider the projection problem

$$P^n_{u_t} g := \arg\min_{v \in \mathcal{T}_{u_t}} \|g - v\|_n^2 \quad \text{with} \quad \|v\|_n^2 := \frac{1}{n} \sum_{i=1}^{n} w(x_i) v(x_i)^2$$

  and with i.i.d. samples $x_i \sim w^{-1}\rho$.

- Under appropriate conditions and with $n \in \mathcal{O}(d \ln(d))$, it can be shown that

$$\mathbb{E}\left[\|g - P^n_{u_t} g\|\right] \lesssim \|g - P_{u_t} g\| .$$

Optimal sampling

- Let $\{b_k\}_{k=1,\ldots,d}$ be an ONB of $\mathcal{T}_{u_t}$ and define $G^n_{kl} := (b_k, b_l)_n$.
- If $\{x_i\}_{i=1,\ldots,n}$ for $n \in \mathcal{O}(d\ln(d))$ are drawn with respect to the density

$$w^{-1}\rho = \frac{1}{d}\sum_{k=1}^{d} b_k^2 \rho \,,$$

  then $\|G^n - I\|_2 \leq \frac{1}{2}$ with high probability.
- If we condition $\{x_i\}_{i=1,\ldots,n}$ to the event $\|G^n - I\|_2 \leq \frac{1}{2}$, then

$$\mathbb{E}\left[\|g - P^n_{u_t} g\|\right] \lesssim \|g - P_{u_t} g\| \,.$$

Algorithm

Setting

Model class

- Let $\mathscr{H}$ be a Hilbert space.
- Consider the model class $\mathscr{M} \subseteq \mathscr{H}$.
- Associate a finite-dimensional subspace $\mathscr{T}_u \subseteq \mathscr{H}$ to every $u \in \mathscr{M}$.
- Let P_u be the $\mathscr{H}$-orthogonal projection onto $\mathscr{T}_u$.

Setting
Optimisation

- Consider the optimisation problem

$$\operatorname*{minimise}_{v \in \mathscr{M}} \mathscr{L}(v)\,.$$

- To update, we project $\nabla \mathscr{L}(u) \in \mathscr{H}$ onto $\mathscr{T}_u$.
- We replace the inaccessible P_u by the n-sample estimate P_u^n .

The Algorithm

1. Compute the gradient

$$g_t := \nabla \mathscr{L}(u_t).$$

2. Compute the local linearisation $\mathscr{T}_{u_t}$ and empirical projection $P^n_{u_t}$.

3. Perform the linear update

$$\bar{u}_{t+1} := u_t - s_t P^n_{u_t} g_t.$$

4. Map $\bar{u}_{t+1} \in \mathscr{H}$ back to $\mathscr{M}$ via the recompression map

$$u_{t+1} := \mathscr{R}_{u_t}(\bar{u}_{t+1})$$

The Algorithm

1. Compute the gradient

$$g_t := \nabla \mathscr{L}(u_t).$$

2. Compute the local linearisation $\mathscr{T}_{u_t}$ and empirical projection $P^n_{u_t}$.
3. Perform the linear update

$$\bar{u}_{t+1} := u_t - s_t P^n_{u_t} g_t.$$

4. Map $\bar{u}_{t+1} \in \mathscr{H}$ back to $\mathscr{M}$ via the recompression map

$\bar{u}_{t+1}$ may not lie in $\mathscr{M}$

$$u_{t+1} := \mathscr{R}_{u_t}(\bar{u}_{t+1})$$

Assumptions

I write $X_t = X_{u_t}$ for any family $\{X_u\}_{u \in \mathscr{M}}$.

Assumption 0

Projection properties

- The random iterates u_t induce the filtration $\mathscr{F}_t = \sigma(u_t, \mathscr{F}_{t-1})$.
- For all $t > 0$ it holds that
  - P_t^n is independent of $\mathscr{F}_t$
  - P_t^n is an unbiased estimator of P_t $\left(\mathbb{E}\left[P_t^n\right] = P_t\right)$
  - $\mathbb{E}\left[\|P_t^n g\|^2\right] \leq \frac{V}{n}\|g\|^2.$

Assumption 0

Projection properties

- The random iterates u_t induce the filtration $\mathscr{F}_t = \sigma(u_t, \mathscr{F}_{t-1})$.
- For all $t > 0$ it holds that
  - P_t^n is independent of $\mathscr{F}_t$
  - P_t^n is an unbiased estimator of P_t **The results hold more generally.**
  - $\mathbb{E}\left[\|P_t^n g\|^2\right] \leq \frac{V}{n}\|g\|^2.$

Assumption 1

Boundedness from below

- $\mathscr{L}_{\min,\mathscr{M}} := \inf_{v \in \mathscr{M}} \mathscr{L}(v)$ is finite.

Assumption 2

L-smoothness

- There exists $L > 0$ such that for all $u, g \in \mathscr{H}$

$$\mathscr{L}(u + g) \leq \mathscr{L}(u) + (\nabla \mathscr{L}(u), g) + \frac{L}{2} \|g\|^2 .$$

Assumption 2

L-smoothness

- There exists $L > 0$ such that for all $u, g \in \mathscr{H}$

$$\mathscr{L}(u+g) \leq \mathscr{L}(u) + (\nabla \mathscr{L}(u), g) + \frac{L}{2}\|g\|^2 .$$

- **There exist local quadratic majorisers.**

Assumption 3

λ-Polyak–Łojasiewicz on $\mathscr{M}$

- $\mathscr{L}_{\min,\mathscr{M}} := \inf_{v \in \mathscr{M}} \mathscr{L}(v)$ is finite.
- There exists $\lambda > 0$ such that for all $u \in \mathscr{M}$

$$\|P_u \nabla \mathscr{L}(u)\|^2 \geq 2\lambda(\mathscr{L}(u) - \mathscr{L}_{\min,\mathscr{M}}) .$$

Assumption 3

λ-Polyak–Łojasiewicz on $\mathscr{M}$

- $\mathscr{L}_{\min,\mathscr{M}} := \inf_{v \in \mathscr{M}} \mathscr{L}(v)$ is finite.
- There exists $\lambda > 0$ such that for all $u \in \mathscr{M}$

$$\|P_u \nabla \mathscr{L}(u)\|^2 \geq 2\lambda(\mathscr{L}(u) - \mathscr{L}_{\min,\mathscr{M}}) \, .$$

- **There exists a global quadratic majoriser.**

Assumption 3

λ-Polyak–Łojasiewicz on $\mathscr{M}$

- $\mathscr{L}_{\min,\mathscr{M}} := \inf_{v \in \mathscr{M}} \mathscr{L}(v)$ is finite.
- There exists $\lambda > 0$ such that for all $u \in \mathscr{M}$

$$\|P_u \nabla \mathscr{L}(u)\|^2 \geq 2\lambda(\mathscr{L}(u) - \mathscr{L}_{\min,\mathscr{M}}) \,.$$

- There exists a global quadratic majoriser. Weaker than strong convexity!

Assumption 3

λ-Polyak–Łojasiewicz on $\mathscr{M}$

- $\mathscr{L}_{\min,\mathscr{M}} := \inf_{v \in \mathscr{M}} \mathscr{L}(v)$ is finite.
- There exists $\lambda > 0$ such that for all $u \in \mathscr{M}$

$$\|P_u \nabla \mathscr{L}(u)\|^2 \geq 2\lambda(\mathscr{L}(u) - \mathscr{L}_{\min,\mathscr{M}}) .$$

- There exists a global quadratic majoriser.
- Stronger than λ-PŁ: Depends not only on $\mathscr{L}$ but also on $\mathscr{M}$.
- (Probably) sufficient condition: PŁ + $\mathscr{M}$ is convex.

Assumption 3

λ-Polyak–Łojasiewicz on $\mathscr{M}$

- $\mathscr{L}_{\min,\mathscr{M}} := \inf_{v \in \mathscr{M}} \mathscr{L}(v)$ is finite.
- There exists $\lambda > 0$ such that for all $u \in \mathscr{M}$

$$\|P_u \nabla \mathscr{L}(u)\|^2 \geq 2\lambda(\mathscr{L}(u) - \mathscr{L}_{\min,\mathscr{M}}) .$$

- There exists a global quadratic majoriser.
- Stronger than λ-PŁ: Depends not only on $\mathscr{L}$ but also on $\mathscr{M}$.
- (Probably) sufficient condition: PŁ + $\mathscr{M}$ is convex.
- Only for simplicity!

Assumption 4

C-controlled retraction error

- There exists a constant $C \geq 0$ such that for all $u \in \mathscr{M}$ and $g \in \mathscr{T}_u$

$$\mathscr{L}(\mathscr{R}_u(u+g)) \leq \mathscr{L}(u+g) + \frac{C}{2}\|g\|^2 .$$

Assumption 4

C-controlled retraction error

- There exists a constant $C \geq 0$ such that for all $u \in \mathscr{M}$ and $g \in \mathscr{T}_u$

$$\mathscr{L}(\mathscr{R}_u(u+g)) \leq \mathscr{L}(u+g) + \frac{C}{2}\|g\|^2 .$$

- **The retraction error is of higher order than the update!**

Assumption 4

C-controlled retraction error

- There exists a constant $C \geq 0$ such that for all $u \in \mathscr{M}$ and $g \in \mathscr{T}_u$

$$\mathscr{L}(\mathscr{R}_u(u+g)) \leq \mathscr{L}(u+g) + \frac{C}{2}\|g\|^2 .$$

- **The retraction error is of higher order than the update!**
- **Not easy to guarantee!**
  - **Compact manifolds with bounded curvature**
  - **Low-rank manifolds with rank-adaptive retraction $\mathscr{R}_u$**

Assumption 4

C-controlled retraction error

- There exists a constant $C \geq 0$ such that for all $u \in \mathscr{M}$ and $g \in \mathscr{T}_u$

$$\mathscr{L}(\mathscr{R}_u(u+g)) \leq \mathscr{L}(u+g) + \frac{C}{2}\|g\|^2 .$$

- The retraction error is of higher order than the update!
- Not easy to guarantee!
  - Compact manifolds with bounded curvature
  - Low-rank manifolds with rank-adaptive retraction $\mathscr{R}_u$
- Only for simplicity?

Results

Descent

under L-smoothness

- Assume $\mathscr{L}$ is L-smooth and $\mathscr{R}$ has C-controlled error. Define

$$c := \frac{L+C}{2} \qquad \text{and} \qquad \sigma_t := s_t - s_t^2 c\left(1 + \frac{V-1}{n}\right).$$

- Then

$$\mathbb{E}\left[\mathscr{L}(u_{t+1}) \mid \mathscr{F}_t\right] \leq \mathscr{L}(u_t) - \sigma_t \|P_t g_t\|^2 + s_t^2 \frac{cV}{n} \|(I - P_t) g_t\|^2 .$$

Convergence under L-smoothness

$$\mathbb{E}\left[\mathscr{L}(u_{t+1}) \mid \mathscr{F}_t\right] \leq \mathscr{L}(u_t) - \sigma_t \|P_t g_t\|^2 + s_t^2 \frac{cV}{n} \|(I - P_t) g_t\|^2$$

- Classical results imply almost sure convergence.
- But s_t must depend on $\|(I - P_t) g_t\|$.
- If the optimum $u^\star \notin \mathscr{M}$, then

$$\|(I - P_t) g_t\| \geq c > 0\,.$$

- This requires $s_t \overset{!}{\to} 0$ to ensure convergence.

Convergence under L-smoothness

Best-case setting

$$\mathbb{E}\left[\mathcal{L}(u_{t+1}) \mid \mathcal{F}_t\right] \le \mathcal{L}(u_t) - \sigma_t\|P_t g_t\|^2 + s_t^2 \frac{cV}{n}\|(I - P_t)g_t\|^2$$

- Assume $\|(I - P_t)g_t\| = 0$.
- Then s_t can be constant and almost surely

$$\min_{t=1,\ldots,\tau} \|P_t g_t\|^2 \in \mathcal{O}(\tau^{\varepsilon - 1})\,.$$

Convergence under L-smoothness

Best-case setting

$$\mathbb{E}\left[\mathcal{L}(u_{t+1}) \mid \mathcal{F}_t\right] \leq \mathcal{L}(u_t) - \sigma_t \|P_t g_t\|^2 + s_t^2 \frac{cV}{n} \|(I - P_t) g_t\|^2$$

- Assume $\|(I - P_t)g_t\| = 0$.
- Then s_t can be constant and almost surely

$$\min_{t=1,\ldots,\tau} \|P_t g_t\|^2 \in \mathcal{O}(\tau^{\varepsilon-1}) \, .$$

This is (almost) the convergence rate of deterministic GD.

Convergence under L-smoothness

Worst-case setting

$$\mathbb{E}\left[\mathscr{L}(u_{t+1}) \mid \mathscr{F}_t\right] \leq \mathscr{L}(u_t) - \sigma_t \|P_t g_t\|^2 + s_t^2 \frac{cV}{n} \|(I - P_t) g_t\|^2$$

- Assume $\|(I - P_t)g_t\| \geq c > 0$ with $\|(I - P_t)g_t\| \in \ell^\infty$.
- Then s_t must obey the Robbins—Monro condition $s_t \in \ell^2$ and $s_t \notin \ell^1$.
- For $s_t \propto t^{-\varepsilon - 1/2}$ with $\varepsilon \in (0, \frac{1}{2})$ the convergence rate is

$$\min_{t=1,\ldots,\tau} \|P_t g_t\|^2 \in \mathcal{O}(\tau^{\varepsilon - 1/2}) .$$

Convergence under L-smoothness

Worst-case setting

$$\mathbb{E}\left[\mathscr{L}(u_{t+1}) \mid \mathscr{F}_t\right] \leq \mathscr{L}(u_t) - \sigma_t \|P_t g_t\|^2 + s_t^2 \frac{cV}{n} \|(I - P_t) g_t\|^2$$

- Assume $\|(I - P_t) g_t\| \geq c > 0$ with $\|(I - P_t) g_t\| \in \ell^\infty$.
- Then s_t must obey the Robbins—Monro condition $s_t \in \ell^2$ and $s_t \notin \ell^1$.
- For $s_t \propto t^{-\varepsilon - 1/2}$ with $\varepsilon \in (0, \frac{1}{2})$ the convergence rate is

$$\min_{t=1,\dots,\tau} \|P_t g_t\|^2 \in \mathcal{O}(\tau^{\varepsilon - 1/2}) \,.$$

This is (almost) the convergence rate of classical SGD.

Descent

under λ-PŁ on $\mathscr{M}$

- Assume $\mathscr{L}$ is L-smooth and λ-PŁ on $\mathscr{M}$ and $\mathscr{R}$ has C-controlled error. Define

$$c := \frac{L+C}{2}, \qquad \sigma_t := s_t - s_t^2 c\left(1+\frac{V-1}{n}\right) \quad \text{and} \quad a_t := 1 - 2\lambda\sigma_t\,.$$

- If $a_t \in (0,1)$, it holds

$$\mathbb{E}\left[\mathscr{L}(u_{t+1}) - \mathscr{L}_{\min,\mathscr{M}} \mid \mathscr{F}_t\right] \leq a_t\left(\mathscr{L}(u_t) - \mathscr{L}_{\min,\mathscr{M}}\right) + s_t^2 \frac{cV}{n} \|(I-P_t)g_t\|^2.$$

Convergence under λ-PŁ on $\mathscr{M}$

$$\mathbb{E}\left[\mathscr{L}(u_{t+1}) - \mathscr{L}_{\min,\mathscr{M}} \mid \mathscr{F}_t\right] \le a_t\left(\mathscr{L}(u_t) - \mathscr{L}_{\min,\mathscr{M}}\right) + s_t^2\frac{cV}{n}\|(I-P_t)g_t\|^2$$

- Classical results imply almost sure convergence.
- But s_t must depend on $\|(I-P_t)g_t\|$.
- If the optimum $u^\star \notin \mathscr{M}$, then

$$\|(I-P_t)g_t\| \ge c > 0\,.$$

- This requires $s_t \overset{!}{\to} 0$ to ensure convergence.

Convergence under λ-PŁ on $\mathscr{M}$

Best-case setting

$$\mathbb{E}\left[\mathscr{L}(u_{t+1}) - \mathscr{L}_{\min,\mathscr{M}} \mid \mathscr{F}_t\right] \leq a_t \left(\mathscr{L}(u_t) - \mathscr{L}_{\min,\mathscr{M}}\right) + s_t^2 \frac{cV}{n} \|(I - P_t)g_t\|^2$$

- Assume $\|(I - P_t)g_t\| = 0$.
- Then s_t can be constant such that $a_t \equiv a \in (0,1)$ and almost surely

$$\mathscr{L}(u_t) - \mathscr{L}_{\min,\mathscr{M}} \in \mathcal{O}(a^{(1-\varepsilon)t})\ .$$

Convergence under λ-PŁ on $\mathscr{M}$

Best-case setting

$$\mathbb{E}\left[\mathscr{L}(u_{t+1}) - \mathscr{L}_{\min,\mathscr{M}} \mid \mathscr{F}_t\right] \leq a_t\left(\mathscr{L}(u_t) - \mathscr{L}_{\min,\mathscr{M}}\right) + s_t^2\frac{cV}{n}\|(I - P_t)g_t\|^2$$

- Assume $\|(I - P_t)g_t\| = 0$.
- Then s_t can be constant such that $a_t \equiv a \in (0,1)$ and almost surely

$$\mathscr{L}(u_t) - \mathscr{L}_{\min,\mathscr{M}} \in \mathscr{O}(a^{(1-\varepsilon)t}) \,.$$

This is (almost) the convergence rate of deterministic GD.

Convergence under λ-PŁ on $\mathscr{M}$

Worst-case setting

$$\mathbb{E}\left[\mathscr{L}(u_{t+1}) - \mathscr{L}_{\min,\mathscr{M}} \mid \mathscr{F}_t\right] \leq a_t\left(\mathscr{L}(u_t) - \mathscr{L}_{\min,\mathscr{M}}\right) + s_t^2\frac{cV}{n}\|(I-P_t)g_t\|^2$$

- Assume $\|(I-P_t)g_t\| \geq c > 0$ with $\|(I-P_t)g_t\| \in \ell^\infty$.
- Then s_t must obey the Robbins—Monro condition $s_t \in \ell^2$ and $s_t \notin \ell^1$.
- For $s_t \in \mathcal{O}(t^{\delta-1})$ with $\delta \in (0,\frac{1}{2})$, the convergence rate is

$$\mathscr{L}(u_t) - \mathscr{L}_{\min,\mathscr{M}} \in \mathcal{O}(t^{\epsilon-1}), \quad \epsilon \in (2\delta,1)\,.$$

Convergence under λ-PŁ on $\mathscr{M}$

Worst-case setting

$$\mathbb{E}\left[\mathscr{L}(u_{t+1}) - \mathscr{L}_{\min,\mathscr{M}} \mid \mathscr{F}_t\right] \leq a_t \left(\mathscr{L}(u_t) - \mathscr{L}_{\min,\mathscr{M}}\right) + s_t^2 \frac{cV}{n} \|(I - P_t)g_t\|^2$$

- Assume $\|(I - P_t)g_t\| \geq c > 0$ with $\|(I - P_t)g_t\| \in \ell^\infty$.
- Then s_t must obey the Robbins—Monro condition $s_t \in \ell^2$ and $s_t \notin \ell^1$.
- For $s_t \in \mathcal{O}(t^{\delta-1})$ with $\delta \in (0, \frac{1}{2})$, the convergence rate is

$$\mathscr{L}(u_t) - \mathscr{L}_{\min,\mathscr{M}} \in \mathcal{O}(t^{\epsilon-1}), \quad \epsilon \in (2\delta, 1)\,.$$

This is (almost) the convergence rate of classical SGD.

Summary

	GD	Best-case	Worst-case	SGD *
L-smoothness	$\mathcal{O}(\tau^{-1})$	$\mathcal{O}(\tau^{\varepsilon-1})$	$\mathcal{O}(\tau^{\varepsilon-1/2})$	$\mathcal{O}(\tau^{\varepsilon-1/2})$
λ-PŁ on $\mathcal{M}$	$\mathcal{O}(a^{\tau})$	$\mathcal{O}(a^{(1-\varepsilon)\tau})$	$\mathcal{O}(\tau^{\varepsilon-1})$	$\mathcal{O}(\tau^{\varepsilon-1})$

*for non-linear model classes