Uniform Convergence of Gradients for Non-Convex Learning and Optimization

Dylan Foster, Ayush Sekhari, and Karthik Sridharan

{djfoster, sekhari, sridharan}@cs.cornell.edu

Cornell University

Overview

When does finding stationary points of the empirical loss imply stationary points for the population loss?

When does finding stationary points of the empirical loss imply low excess risk for the population loss?

Background:

- **Problem Setup.** We aim to solve:

$$\arg\min_{w\in\mathcal{W}} L_{\mathcal{D}}(w) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ell(w; x, y)$$

  - $w \in \mathcal{W} \subseteq \mathbb{R}^d$ is the parameter vector.
  - $\mathcal{D}$ is an unknown probability distribution over the instance space $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^d \times \mathbb{R}$.
  - The loss $\ell(w; x, y)$ is a **potentially non-convex** function of w.
  - Learner gets i.i.d. samples $(x_i, y_i)_{i=1}^n \sim \mathcal{D}$ but **does not observe $\mathcal{D}$ directly**.
  - Learner's performance is quantified by the *excess risk* $L_{\mathcal{D}}(\hat{w}) - L^\star$, where $L^\star = \inf_{w\in\mathcal{W}} L_{\mathcal{D}}(w)$.
- **Gradient Dominance Condition.**

Definition: Gradient Dominance condition

The population risk $L_{\mathcal{D}}$ satisfies the (global) (α, μ)-Gradient Dominance condition with respect to a norm $\|\cdot\|$ if there are constants $\mu > 0$, $\alpha \in [1, 2]$ such that

$$L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^\star) \leq \mu \|\nabla L_{\mathcal{D}}(w)\|^\alpha \quad \forall w \in \mathcal{W},$$

where $w^\star \in \arg\min_{w\in\mathcal{W}} L_{\mathcal{D}}(w)$ is a population minimizer.

Gradient Dominance condition is closely related to **Kurdyka-Łojasiewicz (KŁ)** and **Polyak-Łojasiewicz (PŁ)** inequalities (case $\alpha = 2$).

Contributions

1. **Linear models w/ gradient dominance**: *Any algorithm* that finds stationary points of empirical loss has low excess risk.
   - **Consequence:** Turn learn, suffices to use *any* first order algorithms (gradient descent, SGD, SVRG, SCSG, ...) that finds $\hat{w}$ such that

$$\left\|\nabla \hat{L}_n(\hat{w})\right\| \leq \epsilon.$$

   - Optimal rates both in high- (possibly infinite) dimensional regime and low-dimensional regime.
   - Only need to assume gradient dominance on the population loss.
2. **Non-Smooth Models**: **Dimension-dependent** lower bound. Can circumvent using new margin assumption.

Theorem: Empirical Stationary Point Implies Low Excess Risk

With probability at least $1-\delta$ over the draw of data $(x, y)_{1:n}$, for any algorithm:

- **Smooth high-dimensional setup** - For β-smooth norm $\|\cdot\|$:

$$L_{\mathcal{D}}(\hat{w}^{\mathrm{alg}}) - L^\star \leq \mu_h \cdot \left\|\nabla \hat{L}_n(\hat{w}^{\mathrm{alg}})\right\| + \frac{C_h}{\sqrt{n}}.$$

- **Low-dimensional ℓ_2/ℓ_2 setup** - For $\|\cdot\| = \ell_2$:

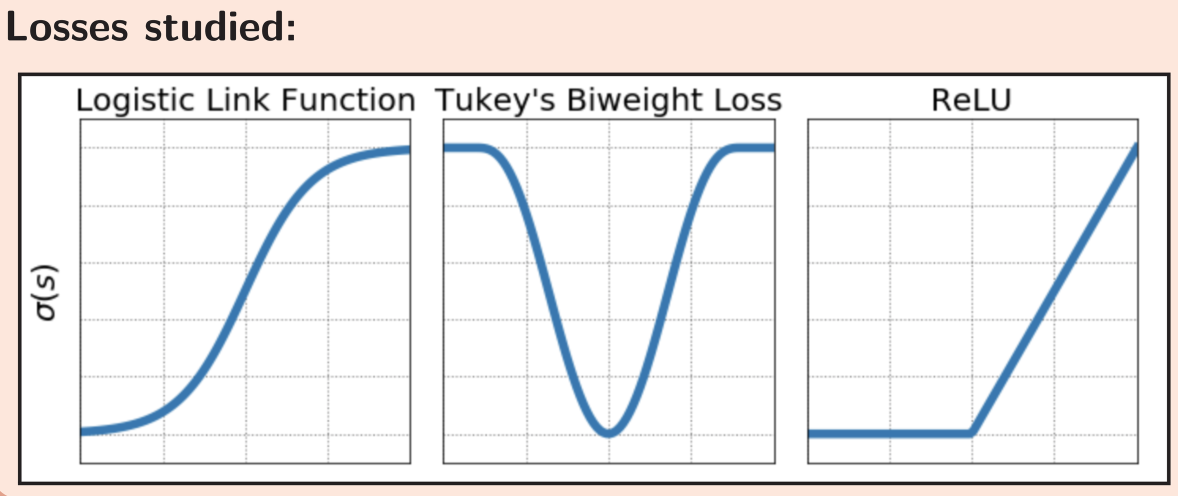
$$L_{\mathcal{D}}(\hat{w}^{\mathrm{alg}}) - L^\star \leq \frac{1}{\lambda_{\min}(\Sigma)}\left(\mu_l \cdot \left\|\nabla \hat{L}_n(\hat{w}^{\mathrm{alg}})\right\|_2^2 + \frac{C_l}{n}\right).$$

- **Sparse ℓ_∞/ℓ_1 setup** - For $\|\cdot\| = \ell_\infty$ and $\|w^\star\|_1 = B$:

$$L_{\mathcal{D}}(\hat{w}^{\mathrm{alg}}) - L^\star \leq \frac{\|w^\star\|_0}{\psi_{\min}(\Sigma)}\left(\mu_s \cdot \left\|\nabla \hat{L}_n(\hat{w}^{\mathrm{alg}})\right\|_\infty^2 + \frac{C_s}{n}\right),$$

where $C_h/C_l/C_s$ and $\mu_h/\mu_l/\mu_s$ are dimension free but problem dependent constants.

Losses studied:



Tools

Dimension-Free Gradient Uniform Convergence.

$$\sup_{w\in\mathcal{W}} \left\|\nabla \hat{L}_n(w) - \nabla L_{\mathcal{D}}(w)\right\| \leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right)$$

via **Normed Rademacher Complexity**

$$\mathfrak{R}_{\|\cdot\|}(\mathcal{F};\ z_{1:n}) := \mathop{\mathbb{E}}_{\epsilon} \sup_{f\in\mathcal{F}} \left\|\sum_{t=1}^n \epsilon_t f(z_t)\right\|,$$

where $f \in \mathcal{F} : \mathbb{R}^d \mapsto \mathbb{R}^k$ and ϵ_i is ± 1 with probability half.

Lemma

For any $\delta > 0$, with probability at least $1-\delta$ over the examples $x_{1:n}, y_{1:n}$,

$$\mathop{\mathbb{E}} \sup_{w\in\mathcal{W}} \left\|\nabla \hat{L}_n(w) - \nabla L_{\mathcal{D}}(w)\right\| \leq \frac{4}{n}\mathfrak{R}_{\|\cdot\|}(\nabla\ell\circ\mathcal{W};\ (x,y)_{1:n}) + c\frac{\log\left(\frac{1}{\delta}\right)}{n}.$$

Vector-Valued Rademacher Complexity

$$\vec{\mathfrak{R}}(\mathcal{G};\ z_{1:n}) := \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \sup_{g\in\mathcal{G}} \sum_{t=1}^n \langle \boldsymbol{\epsilon}_t, g(z_t)\rangle$$

where $g \in \mathcal{G} : \mathbb{R}^d \mapsto \mathbb{R}^k$ and $\boldsymbol{\epsilon}_i \in \mathbb{R}^k$ is a vector of independent Rademacher variables.

Chain Rule for Rademacher Complexity

Theorem

For $G_t : \mathbb{R}^K \to \mathbb{R}$ and $F_t : \mathbb{R}^d \to \mathbb{R}^K$ with $\|\nabla G_t\|_2 \leq L_G$ and $\sqrt{\sum_{k=1}^K \|\nabla F_{t,k}(w)\|^2} \leq L_F$, the *normed Rademacher complexity* of the composition is bounded as

$$\frac{1}{2}\mathop{\mathbb{E}}_{\epsilon} \sup_{w\in\mathcal{W}} \left\|\sum_{t=1}^n \epsilon_t \nabla(G_t(F_t(w)))\right\| \leq$$

$$L_F \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \sup_{w\in\mathcal{W}} \sum_{t=1}^n \langle \boldsymbol{\epsilon}_t, \nabla G_t(F_t(w))\rangle + L_G \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}} \sup_{w\in\mathcal{W}} \left\|\sum_{t=1}^n \nabla F_t(w)\boldsymbol{\epsilon}_t\right\|.$$

Applications

Generalized Linear Model

Setup: $\ell(w;\ x, y) = (\sigma(\langle w, x\rangle) - y)^2$, where,

1. $\sigma : \mathbb{R} \to [0, 1]$ is a smooth, potentially non-convex link function.
2. $\mathcal{X} \subseteq \{x \in \mathbb{R}^d \mid \|x\| \leq R\}$ and $\mathcal{W} \subseteq \{w \in \mathbb{R}^d \mid \|w\|_\star \leq B\}$; $\mathcal{Y} = \{0, 1\}$.

Example: logistic link function

$$\sigma(s) = (1 + e^{-s})^{-1}.$$

Assumptions: (Generalized Linear Model Regularity) Let $\mathcal{S} = [-BR, BR]$.

(a) $\exists C_\sigma \geq 1$ s.t. $\max\{\sigma'(s), \sigma''(s)\} \leq C_\sigma$ for all $s \in \mathcal{S}$.

(b) $\exists c_\sigma > 0$ s.t. $\sigma'(s) \geq c_\sigma$ for all $s \in \mathcal{S}$.

(c) $\mathbb{E}[y \mid x] = \sigma(\langle w^\star, x\rangle)$ for some $w^\star \in \mathcal{W}$.

Robust Regression

Setup: $\ell(w;\ x, y) = \rho(\langle w, x\rangle - y)$, where,

1. $\rho : \mathbb{R} \to \mathbb{R}$ is a smooth, potentially non-convex function.
2. $\mathcal{X} \subseteq \{x \in \mathbb{R}^d \mid \|x\| \leq R\}$ and $\mathcal{W} \subseteq \{w \in \mathbb{R}^d \mid \|w\|_\star \leq B\}$; $\mathcal{Y} = \mathbb{R}$.

Example: Tukey's biweight loss (for some fixed t_0)

$$\rho(t) = \begin{cases} 1 - (1 - (t/t_0)^2)^3 & |t| \leq t_0, \\ 1 & |t| \geq t_0. \end{cases}$$

Assumptions: (Robust Regression Regularity) Let $\mathcal{S} = [-(BR+Y), (BR+Y)]$.

(a) $\exists C_\rho \geq 1$ s.t. $\max\{\rho'(s), \rho''(s)\} \leq C_\rho$ for all $s \in \mathcal{S}$.

(b) ρ' is odd with $\rho'(s) > 0$ for all $s > 0$ and $h(s) := \mathbb{E}_\zeta[\rho'(s + \zeta)]$ has $h'(0) > c_\rho$.

(c) $\exists w^\star \in \mathcal{W}$ such that $y = \langle w^\star, x\rangle + \zeta$, and ζ is symmetric zero-mean given x.

Sample Complexity Results from Optimization

1. **Algorithms that find stationary points of $\hat{L}_n(\cdot)$ have low excess risk.**

Meta-Algorithm: Learning by Finding Stationary Points

Consider the following meta-algorithm:

(a) Gather $n = \tilde{O}(\frac{1}{\varepsilon^2} \wedge \frac{d}{\varepsilon})$ samples $(x_i, y_i)_{i=1}^n$, and let $\hat{L}_n^\lambda(w) = \hat{L}_n(w) + \frac{\lambda}{2}\|w\|_2^2$.

(b) Find $\hat{w}^{\mathrm{alg}} \in \mathcal{W}$ such that $\nabla \hat{L}_n^\lambda(\hat{w}^{\mathrm{alg}}) = 0$, which is guaranteed to exists.

Then, for appropriate λ, with probability at-least $1 - O(1/n)$,

$$L_{\mathcal{D}}(\hat{w}^{\mathrm{alg}}) - L^\star \leq \varepsilon.$$

* Using any black-box stationary point finding algorithm (Gradient Descent, SGD, SVRG, etc).

2. **Optimal sample complexity.** Our sample complexity $n = \tilde{O}(\frac{1}{\varepsilon^2} \wedge \frac{d}{\varepsilon})$ is optimal up to problem dependent constants [see eg. Tsybakov 2008].

Model	Algorithm	Sample Complexity	
		Norm-based/Infinite dim.	Low-dim.
Generalized Linear	Ours	$O(\varepsilon^{-2})$	$O(d\varepsilon^{-1})$
	Mei et al. (2016) - Theorem 4	n/a	$O(d\varepsilon^{-1})$
	Kakade et al. (2011) - GLMtron	$O(\varepsilon^{-2})$	n/a
Robust Regression	Ours	$O(\varepsilon^{-2})$	$O(d\varepsilon^{-1})$
	Mei et al. (2016) - Theorem 6	n/a	$O(d\varepsilon^{-1})$

3. **Comparison to Mei et al. (2016).** Our rates match Mei et al. (2016) for GLM and RR under the natural setting of $R = \sqrt{d}$, but avoid explicit dependence on dimension d.

Bound Analyzed	Mei et al. (2016)	Our Results
Uniform Convergence - $\sup_{w\in\mathcal{W}} \left\|\nabla L_{\mathcal{D}}(w) - \nabla\hat{L}_n(w)\right\|_2$	$O\left(R\sqrt{\frac{d}{n}}\right)$	$O\left(R^2\sqrt{\frac{1}{n}}\right)$
Parameter Convergence - $\left\|\hat{w}^{\mathrm{alg}} - w^\star\right\|_2$	$O\left(\frac{R}{\lambda_{\min}(\Sigma)}\sqrt{\frac{d}{n}}\right)$	$O\left(\frac{R^2}{\lambda_{\min}(\Sigma)\sqrt{n}}\right)$

*Song Mei, Yu Bai, and Andrea Montanari, *The Landscape of Empirical Risk for Nonconvex Losses.*, 2016

Non-Smooth Models

Suppose the loss is non-smooth, When does finding stationary points of $\hat{L}_n(\cdot)$ imply stationary points of $L_{\mathcal{D}}(\cdot)$?

Case Study: Single ReLU

Setup: $\ell(w;\ x, y) = \mathrm{ReLU}(-\langle w, x\rangle \cdot y)$, where,

$$\mathcal{X} \subseteq \{x \in \mathbb{R}^d \mid \|x\|_2 \leq 1\},\ \mathcal{Y} = \{-1, +1\},\ \mathcal{W} \subseteq \{w \in \mathbb{R}^d \mid \|w\|_2 \leq 1\}.$$

Result 1: Dimension-dependent lower bound.

Theorem: Dimension Dependent Lower Bound

For all $n \in \mathbb{N}$ there exist a sequence of instances $x_{1:n}, y_{1:n}$ such that

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{w\in\mathcal{W}} \left\|\sum_{t=1}^n \epsilon_t \nabla\ell(w;\ x_t, y_t)\right\|_2 = \Omega\left(\sqrt{dn} \wedge n\right).$$

Result 2: Circumventing the lower bound.

Parameters $w \in \mathcal{W}$ satisying an additional margin-type assumption enjoy dimension free uniform convergence of gradients (parameterized by ϕ).

Definition: ϕ-soft-margin condition

Given an increasing function $\phi : [0, 1] \to [0, 1]$, and a distribution P over the support $\mathcal{X}$, the weight vector $w \in \mathcal{W}$ satisfies **ϕ-soft-margin condition with respect to P** if :

$$\mathop{\mathbb{E}}_{x\sim P}\left[\mathbb{1}\left\{\frac{|\langle w, x\rangle|}{\|w\|_2\|x\|_2} \leq \gamma\right\}\right] \leq \phi(\gamma) \qquad \forall\gamma$$

Theorem: Uniform Convergence of Gradients

For a fixed ϕ, define $\mathcal{W}(\phi; \hat{\mathcal{D}}_n)$ as the subset of $\mathcal{W}$ which satisfy ϕ-soft-margin assumption w.r.t. the empirical data distribution, i.e.:

$$\mathcal{W}(\phi; \hat{\mathcal{D}}_n) = \left\{w \in \mathcal{W}\ :\ \begin{array}{c} w \text{ satisfies } \phi\text{-soft-margin condition} \\ \text{w.r.t. empirical data distribution } \hat{\mathcal{D}}_n \end{array}\right\}$$

Then, with probability at least $1-\delta$ over the choice of samples,

$$\sup_{w\in\mathcal{W}(\phi,\hat{\mathcal{D}}_n)} \left\|\nabla L_{\mathcal{D}}(w) - \nabla \hat{L}_n(w)\right\|_2$$

$$\leq \tilde{O}\left(\inf_{\gamma>0}\left\{\sqrt{\phi(4\gamma)} + \frac{1}{\gamma}\sqrt{\frac{\log(1/\delta)}{n}}\right\} + \frac{1}{\sqrt{\gamma}n^{1/4}}\right).$$

Example: When $\phi(\gamma) = \gamma^{\frac{1}{2}}$, the above yields bound of $O(n^{-1/12})$.