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:

  \[
  \arg \min_{w \in W} L_{\text{emp}}(w) = \mathbb{E}_{(x,y) \sim D}(f(w, x, y))
  \]

  where \(w \in W \subseteq \mathbb{R}^d\) is the parameter vector.

  - \(D\) is an unknown probability distribution over the instance space \(X \times Y \subseteq \mathbb{R}^d \times \mathbb{R}\).

  - The loss \(f(w, x, y)\) is a potentially non-convex function of \(w\).

  - Learner gets iid samples \((x,y)\) and \(D\) but does not observe \(D\) directly.

  - Learner’s performance is quantified by the excess risk \(L_{\text{emp}}(w) - L^*\), where \(L^* = \inf_{w \in W} L_{\text{emp}}(w)\).

  - Gradient Dominance Condition.

Definition: Gradient Dominance condition

The population risk \(L(w)\) satisfies the (global) \((\rho, \mu)\)-Dominance condition with respect to a norm \(\|\cdot\|\) if there are constants \(\rho > 0, \mu \in [0, 2]\) such that

\[
L_{\text{emp}}(w) - L_{\text{emp}}(w^*) \leq \rho \|\nabla L_{\text{emp}}(w^*)\|_{\text{alg}} + \mu \|w - w^*\|
\]

Gradient Dominance condition is closely related to Kurdyka-Łojasiewicz (KL) and Polyak-Łojasiewicz (PL) inequalities (case \(\mu = 2\)).

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

     \[
     \|\nabla L_{\text{emp}}(\hat{w})\|_{\text{alg}} \leq \varepsilon
     \]

   - Optimal rates both in high- (possibly infinite) dimensional regime and low-dimensional regime.

   - Only need to assume gradient dominance on the population loss.


Theorem: Empirical Stationary Point Satisfies Low Excess Risk

With probability at least \(1 - \delta\) over the draw of data \((x,y)\), for any algorithm,\n
- Smooth high-dimensional setup - For \(\beta\)-smooth norm \(\|\cdot\|\):

  \[
  L_{\text{emp}}(w^*) - L^* \leq \rho \|\nabla L_{\text{emp}}(w^*)\|_{\text{alg}} + \rho \frac{\beta}{\delta}
  \]

- Low-dimensional \(L_2\) / \(L_2-L_2\) setup - For \(\|\cdot\|_2 \leq L_2 \leq \|\cdot\|_{\text{alg}}\):

  \[
  L_{\text{emp}}(w^*) - L^* \leq \frac{\rho}{\sqrt{\delta}} \left( \|\nabla L_{\text{emp}}(w^*)\|_{\text{alg}} + L_2 \right)
  \]

- Sparse / \(L_1\) / \(L_1-L_1\) setup - For \(\|\cdot\|_1 \leq L_1 \leq \|\cdot\|_{\text{alg}}\):

  \[
  L_{\text{emp}}(w^*) - L^* \leq \frac{\rho}{\sqrt{\delta}} \left( \|\nabla L_{\text{emp}}(w^*)\|_{\text{alg}} + L_1 \right)
  \]

where \(C_{\beta}/C_{\beta}\) and \(\mu_{\beta}/\mu_{\beta}\) are dimension free but problem dependent constants.

Losses studied:

- Logistic Link Function
- Tukey’s Biweight Loss
- RELU

Tools

Dimension-Free Gradient Uniform Convergence.

\[
\sup_{w \in W} \|\nabla L_{\text{emp}}(w) - \nabla L_{\text{emp}}(\hat{w})\| \leq \tilde{O} \left( \frac{1}{\sqrt{n}} \right)
\]

via Normed Rademacher Complexity

\[
R_{\text{Rad}}(B_{\text{alg}}) = \mathbb{E}_{\epsilon \sim \mathcal{R}_n} \left[ \sup_{w \in B_{\text{alg}}} \sum_{i=1}^n \epsilon_i f(w^i) \right]
\]

where \(f \in F: \mathbb{R}^d \to \mathbb{R}^d\) and \(\epsilon_i\) is iid with probability half.

Lemma

For any \(\delta > 0\), with probability at least \(1 - \delta\) over the examples \(x_1,\ldots, x_n\),

\[
\mathbb{E}_{x \sim D} \left[ \|\nabla L(w^*) - \nabla L_{\text{emp}}(w)\| \right] \leq \frac{L^*}{n} \mathbb{E}_{x \sim D} \left[ \|\nabla f(x, y)\| + \log \left( \frac{1}{\delta} \right) \right]
\]

Vector-Valued Rademacher Complexity

\[
\mathbb{E} \left[ \sup_{w \in W} \sum_{i=1}^n \epsilon_i f(w^i) \right] \leq \mathbb{E} \left[ \|\nabla \Phi(w) \|_{\text{alg}} \right]
\]

where \(\Phi: W \to \mathbb{R}^d\) is a vector of independent random variables.

Sample Complexity Results from Optimization

1. Algorithms that find stationary points of \(L_{\text{emp}}(w)\) have low excess risk.

   Meta-Algorithm: Learning by Finding Stationary Points

   Consider the following meta-algorithm:

   - Gather \(n = O(1/\delta + 1/\epsilon)\) samples \((x_i, y_i)\) and let \(L_{\text{emp}}(w) = L_{\text{emp}}(w) + \epsilon x_i^T f(w)\).

   - Find \(\hat{w}\) such that \(\nabla L_{\text{emp}}(\hat{w}) = 0\), which is guaranteed to exist.

   Then, for appropriate \(\lambda\) with probability at least \(1 - O(1/n)\),

   \[
   L_{\text{emp}}(\hat{w}) - L^* \leq \epsilon
   \]

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

2. Optimal sample complexity. Our sample complexity \(n = O(1/\delta + 1/\epsilon)\) is optimal up to problem dependent constants [see e.g. Tsybakov 2008].

Non-Smooth Models

Suppose the loss is non-smooth, When does finding stationary points of \(L_{\text{emp}}(w)\) imply stationary points of \(L_{\text{emp}}(w)\)?

Case Study: Finding Single ReLU

Setup: \(f(w, x, y) = \text{ReLU}(\langle w, x \rangle - y)\), where,

\[
X \subseteq \{ x \in \mathbb{R}^d \mid \|x\|_2 \leq 1 \}, \quad Y = \{-1, 1\}, \quad W \subseteq \{ w \in \mathbb{R}^d \mid \|w\|_2 \leq 1 \}
\]

Result 1: Dimension-dependent lower bound.

Theorem: Dimension Dependent Lower Bound

For all \(n \in \mathbb{N}\) there exist a sequence of instances \(x_1,\ldots, x_n\), such that

\[
\mathbb{E} \left[ \sup_{w \in W} \sum_{i=1}^n \epsilon_i \left[ f(x_i, y_i, w^i) \right] \right] \geq \Omega(1/\sqrt{n})
\]

Result 2: Circumventing the lower bound.

Parameters \(w \in W\) satisfying an additional margin-type assumption enjoy dimension free uniform convergence of gradients (parameterized by \(\delta\)).

Definition: \(\delta\)-margin condition

Given an increasing function \(\phi: [0, 1] \to [0, 1]\), and a distribution \(P\) over the support \(X\), the weight vector \(v \in W\) satisfies \(\delta\)-margin condition with respect to \(P\), if

\[
\mathbb{E} \left[ \sup_{w \in W} \left( \langle w, x \rangle + \phi(\|w\|_2) \right) \right] \leq \delta \quad \forall \langle x, y \rangle \in X \times Y
\]

Theorem: Uniform Convergence of Gradients

For a fixed \(\phi\), define \(W_\phi(W)\) as the subset of \(W\) which satisfy \(\delta\)-margin assumption w.r.t. the empirical data distribution, i.e.

\[
W_{\phi}(W) = \left\{ w \in W \mid \text{satisfies } \delta\text{-margin condition w.r.t. empirical data distribution } P_n \right\}
\]

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

\[
\sup_{w \in W_{\phi}(W)} \left\| \nabla L_{\text{emp}}(w) - \nabla L_{\text{emp}}(\hat{w}) \right\| \leq \tilde{O} \left( \frac{1}{\sqrt{n}} \right)
\]

Example: When \(\phi(\cdot) = \gamma^2\), the above holds for \(O(n^{-1/2})\).