Connecting Optimization and Regularization Paths

Arun Sai Suggala, Adarsh Prasad, Pradeep Ravikumar

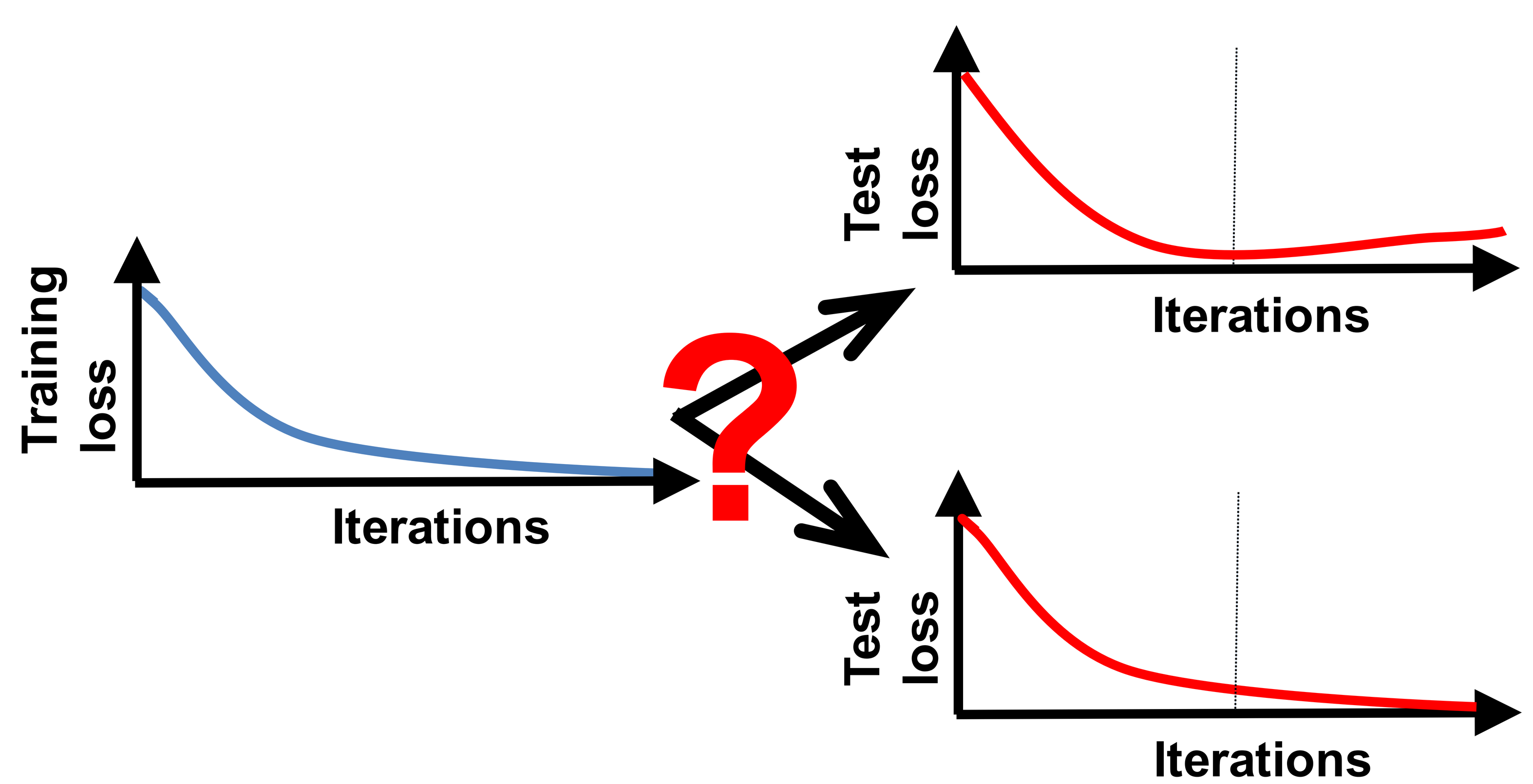
Carnegie Mellon University

Contributions

- Study the **implicit** regularization properties of optimization techniques.
  - **Explicitly** connect their optimization paths to the regularization paths of *corresponding* regularized problems.
- **Strongly Convex Losses**: Both the paths are point-wise close to each other.
  - **Consequences**: Obtain excess risk of iterates of GD, early stopping rules for risk minimization.
- **Convex Losses**: The paths need not always lie close to each other.
  - For linear **classification** with convex surrogates, the paths are close to each other.

Motivation and Setup

- Ambiguity in behavior of Test loss vs Iterations



Setup

- Gradient Descent/Flow on $f(\theta)$:

$$\frac{d}{dt}\theta(t) = -\nabla f(\theta(t)), \quad \theta(0) = \theta_0.$$

- Corresponding Regularized Objective:

$$\theta(\nu) = \arg\min_{\theta} f(\theta) + \frac{1}{2\nu}\|\theta - \theta_0\|_2^2.$$

- GD Path: $\{\theta(t)\}_{t=0}^{\infty}$.
- Regularization Path: $\{\theta(\nu)\}_{\nu=0}^{\infty}$.

Strongly Convex Loss

Theorem 1 *Let f be m strongly convex and M smooth and $c = \frac{2m}{m+M}$. Moreover, let the regularization penalty ν and time t be related through the relation $\nu(t) = \frac{1}{cm}(e^{cMt} - 1)$. Then*

$$\|\theta(t) - \theta(\nu(t))\|_2 \le \frac{\|\nabla f(\theta_0)\|_2}{m}\left(e^{-mt} - \frac{c}{e^{cMt} + c - 1}\right)$$

- When $m = M$, both the paths are the same.
- Both the paths are within $O(e^{-mt} - ce^{-cMt})$ of each other
  - Early stopping GD has **regularization** effect.

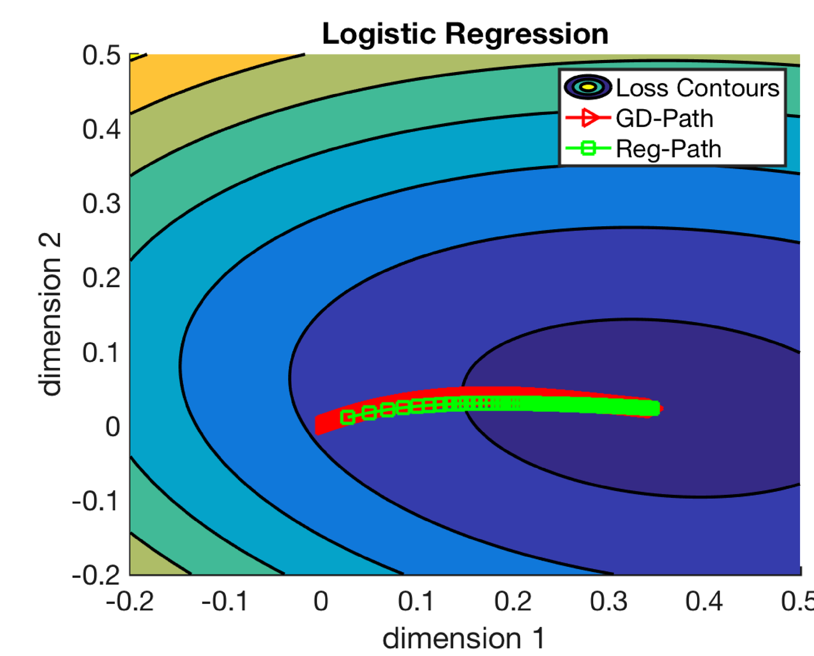


Figure 1: Logistic Regression with inseparable data

Excess Risk of GD Iterates

- $R(\theta), R_n(\theta)$ - population, empirical risks, θ^*- true parameter.

Theorem 2 *For $t \le \frac{1}{cM}\log\left(1 + \frac{cm\|\theta^*\|}{2\|\nabla R_n(\theta^*)\|_2}\right)$, GD iterates $\theta(t)$ satisfy*

$$\|\theta(t) - \theta^*\|_2 \le \frac{\|\nabla R_n(\theta_0)\|_2}{m}\left(e^{-mt} + \frac{c}{1 - c - e^{cMt}}\right) + \frac{3}{c}\frac{e^{-cMt}}{1 - e^{-cMt}}\|\theta^*\|_2.$$

- Roughly speaking, at $t = O\left(\log\left(1 + \frac{m\|\theta^*\|}{2\|\nabla R_n(\theta^*)\|}\right)\right)$ we have

$$\|\theta(t) - \theta^*\|_2 = O\left(\left(e^{-mt} - ce^{-Mt}\right)\|\theta^*\| + \|\nabla R_n(\theta^*)\|\right)$$

Linear Regression - Early Stopping Rule

Corollary 1 *Suppose the covariate vector x has a normal distribution with mean 0 and identity covariance matrix. Then at $t = O\left(\log\left(1 + c_1^2\frac{\|\theta^*\|^2}{\sigma^2}\frac{n}{p}\right)\right)$, the iterate $\theta(t)$ satisfies*

$$\|\theta(t) - \theta^*\|_2^2 \le (1 + \epsilon)\left[\frac{\|\theta^*\|^2}{\|\theta^*\|^2 + \frac{\sigma^2 p}{n}}\right]\frac{\sigma^2 p}{n},$$

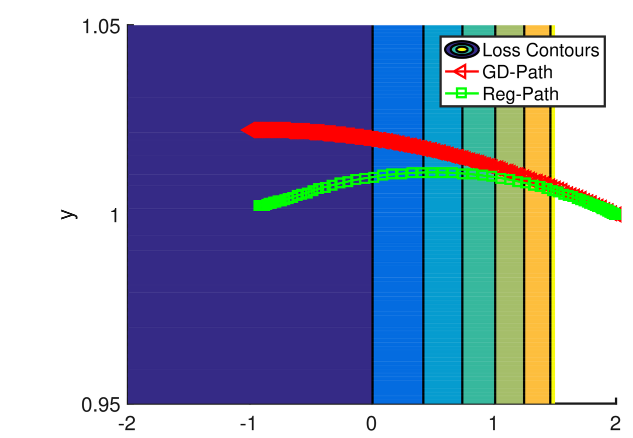
where ϵ is less than 0.1.

Convex Loss

- The paths need not always lie close to each other.
  - Converge to different points
  - Regularization path always converges to **closest** minimizer to initialization point, whereas GD may not.
- Counterexample:

$$f(x, y) = \frac{(x+1)^2}{y + 100}, \text{ for } y > 100, \quad (x_0, y_0) = (2, 1).$$

$$\lim_{t\to\infty}\theta(t) = (-1, 1.02), \quad \lim_{\nu\to\infty}\theta(\nu) = (-1, 1).$$

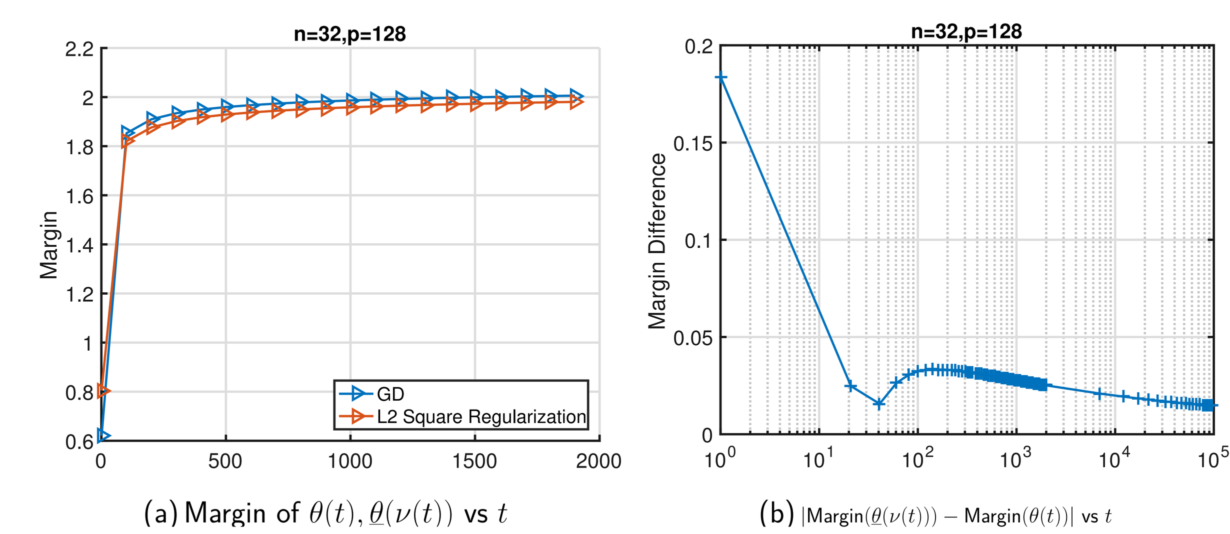


Linear Classification

Theorem 3 *Assume the data D_n is linearly separable. Suppose we use exponential loss to learn a linear classifier. Suppose the regularization parameter ν and time t are related as $\nu(t) = t$. Then for any $t \ge 0$, we have*

$$|Margin(\theta(t)) - Margin(\theta(\nu(t)))| \le O\left(\frac{1}{\log t}\right),$$

where margin of a classifier is the distance of closest point to the decision boundary.



Summary

Table of Connections

Problem	Algorithm	Connected to Regularized Problem?	Metric	$\nu(t)$	Connection
Strongly Convex	Gradient Descent	**Yes**	Parameter Distance	$O(e^{cMt} - 1)$	$O(e^{-mt} - ce^{-cMt})$
Strongly Convex	Mirror Descent	**Yes**	Parameter Distance	$O(e^{cMt/\alpha} - 1)$	$O(e^{-mt/\beta} - ce^{-cMt/\alpha})$
Convex	Gradient Descent	**No**	-	-	-
Classification with exp. loss	Gradient Descent	**Yes**	Margin	$\frac{1}{t}$	$O\left(\frac{1}{\log t}\right)$