The Implicit Regularization of Stochastic Gradient Flow for Least Squares

Alnur Ali[1], Edgar Dobriban[2], and Ryan J. Tibshirani[3]

[1]Stanford University, [2]University of Pennsylvania, [3]Carnegie Mellon University

Outline

Introduction

- Given the sizes of modern data sets, stochastic gradient descent is one of the most widely used optimization algorithms today
  - Computational and statistical properties have been studied for decades (Robbins & Monro, 1951; Fabian, 1968; Ruppert, 1988; Kushner & Yin, 2003; Polyak & Juditsky, 1992; ...)

Introduction

- Given the sizes of modern data sets, stochastic gradient descent is one of the most widely used optimization algorithms today
  - Computational and statistical properties have been studied for decades (Robbins & Monro, 1951; Fabian, 1968; Ruppert, 1988; Kushner & Yin, 2003; Polyak & Juditsky, 1992; ...)

- Recently, lots of interest in implicit regularization

- In particular, a line of work showing (early-stopped) gradient descent is linked to ℓ_2 regularization

Introduction

- Given the sizes of modern data sets, stochastic gradient descent is one of the most widely used optimization algorithms today
  - Computational and statistical properties have been studied for decades (Robbins & Monro, 1951; Fabian, 1968; Ruppert, 1988; Kushner & Yin, 2003; Polyak & Juditsky, 1992; ...)

- Recently, lots of interest in implicit regularization

- In particular, a line of work showing (early-stopped) gradient descent is linked to ℓ_2 regularization

- Interesting, but also computationally convenient

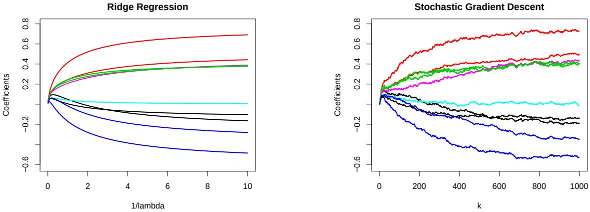
Introduction

- Natural to ask: do the iterates generated by (mini-batch) stochastic gradient descent also possess (implicit) ℓ_2 regularity?

Introduction

- Natural to ask: do the iterates generated by (mini-batch) stochastic gradient descent also possess (implicit) ℓ_2 regularity?

- Why might there be a connection, at all?
  - Compare the paths for least squares regression



- In this paper, we'll focus on least squares regression

Introduction

- Main tool for making the connection: a stochastic differential equation that we call stochastic gradient flow
  - Linked to SGD with a constant step size; more on this later

- We give a bound on the excess risk of stochastic gradient flow at time t, over ridge regression with tuning parameter $\lambda = 1/t$
  - Result(s) hold across the entire optimization path
  - Results do not place strong conditions on the features
  - Proofs are simpler than in discrete-time

Introduction

- Main tool for making the connection: a stochastic differential equation that we call stochastic gradient flow
  - Linked to SGD with a constant step size; more on this later

- We give a bound on the excess risk of stochastic gradient flow at time t, over ridge regression with tuning parameter $\lambda = 1/t$
  - Result(s) hold across the entire optimization path
  - Results do not place strong conditions on the features
  - Proofs are simpler than in discrete-time

- Roughly speaking, the bound decomposes into three parts
  - The variance of ridge regression scaled by a constant less than 1
  - The "price of stochasticity": a term that is non-negative, but vanishes as time grows
  - A term that is tied to the limiting optimization error: this term is zero in the overparametrized regime, but positive otherwise

Outline

Stochastic gradient flow

- We consider the stochastic differential equation

$$d\beta(t) = \underbrace{\frac{1}{n} X^T (y - X\beta(t))\, dt}_{\substack{\text{just the gradient for} \\ \text{least squares regression}}} + \underbrace{Q_\epsilon(\beta(t))^{1/2}\, dW(t),}_{\substack{\text{fluctuations are governed by the} \\ \text{cov. of the stochastic gradients}}} \tag{1}$$

  where $\beta(0) = 0$,

$$Q_\epsilon(\beta) = \epsilon \cdot \mathrm{Cov}_{\mathcal{I}} \left(\frac{1}{m} X_{\mathcal{I}}^T (y_{\mathcal{I}} - X_{\mathcal{I}} \beta) \right)$$

  is the diffusion coefficient, $\mathcal{I} \subseteq \{1, \ldots, n\}$ is a mini-batch, and $\epsilon > 0$ is a (fixed) step size

- We call (1) stochastic gradient flow
  – Has a few nice properties, and bears several connections to SGD with a constant step size; more on this next

Stochastic gradient flow

- Lemma: the Euler discretization of stochastic gradient flow $\tilde{\beta}^{(k)}$, and constant step size SGD $\beta^{(k)}$, share first and second moments, i.e.,

$$\mathbb{E}(\tilde{\beta}^{(k)}) = \mathbb{E}(\beta^{(k)}) \quad \text{and} \quad \mathrm{Cov}(\tilde{\beta}^{(k)}) = \mathrm{Cov}(\beta^{(k)})$$

Stochastic gradient flow

- Lemma: the Euler discretization of stochastic gradient flow $\tilde{\beta}^{(k)}$, and constant step size SGD $\beta^{(k)}$, share first and second moments, i.e.,

$$\mathbb{E}(\tilde{\beta}^{(k)}) = \mathbb{E}(\beta^{(k)}) \quad \text{and} \quad \mathrm{Cov}(\tilde{\beta}^{(k)}) = \mathrm{Cov}(\beta^{(k)})$$

  - Implies the prediction errors match
  - Also, implies any deviation between the first two moments of stochastic gradient flow and SGD must be due to discretization error

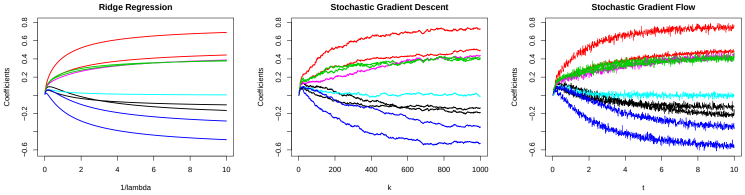
Stochastic gradient flow

- Lemma: the Euler discretization of stochastic gradient flow $\tilde{\beta}^{(k)}$, and constant step size SGD $\beta^{(k)}$, share first and second moments, i.e.,

$$\mathbb{E}(\tilde{\beta}^{(k)}) = \mathbb{E}(\beta^{(k)}) \quad \text{and} \quad \mathrm{Cov}(\tilde{\beta}^{(k)}) = \mathrm{Cov}(\beta^{(k)})$$

  - Implies the prediction errors match
  - Also, implies any deviation between the first two moments of stochastic gradient flow and SGD must be due to discretization error

- Sanity check: revisiting the solution/optimization paths from earlier



Stochastic gradient flow

- A number of works consider instead the constant covariance process,

$$d\beta(t) = \frac{1}{n} X^T (y - X\beta(t))\, dt + \left(\frac{\epsilon}{m} \cdot \hat{\Sigma}\right)^{1/2} dW(t), \tag{2}$$

  where $\hat{\Sigma} = X^T X/n$ (cf. Langevin dynamics)

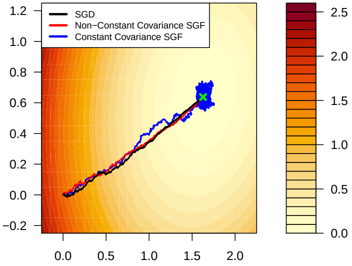
Stochastic gradient flow

- A number of works consider instead the constant covariance process,

$$d\beta(t) = \frac{1}{n} X^T (y - X\beta(t))\, dt + \left(\frac{\epsilon}{m} \cdot \hat{\Sigma}\right)^{1/2} dW(t), \qquad (2)$$

  where $\hat{\Sigma} = X^T X / n$ (cf. Langevin dynamics)

- Turns out (theoretically, empirically) stochastic gradient flow is a more accurate approximation to SGD than (2) is



Outline

Setup

- Assume a standard regression model

$$y = X\beta_0 + \eta, \quad \eta \sim (0, \sigma^2 I)$$

- Fix X; let $s_i, i = 1, \ldots, p$, denote the eigenvalues of $X^T X/n$

Setup

- Assume a standard regression model

$$y = X\beta_0 + \eta, \quad \eta \sim (0, \sigma^2 I)$$

- Fix X; let $s_i, i = 1, \ldots, p$, denote the eigenvalues of $X^T X/n$

- Recall a useful result for (batch) gradient flow (Ali et al., 2018)
  - For least squares regression, gradient flow is

$$\dot{\beta}(t) = \frac{1}{n} X^T (y - X\beta(t))dt, \quad \beta(0) = 0$$

  - Has the solution

$$\hat{\beta}^{\mathrm{gf}}(t) = (X^T X)^+ \big(I - \exp(-tX^T X/n)\big) X^T y$$

Setup

- Assume a standard regression model

$$y = X\beta_0 + \eta, \quad \eta \sim (0, \sigma^2 I)$$

- Fix X; let $s_i, i = 1, \ldots, p$, denote the eigenvalues of $X^T X/n$

- Recall a useful result for (batch) gradient flow (Ali et al., 2018)
  - For least squares regression, gradient flow is

$$\dot{\beta}(t) = \frac{1}{n} X^T (y - X\beta(t))dt, \quad \beta(0) = 0$$

  - Has the solution

$$\hat{\beta}^{\mathrm{gf}}(t) = (X^T X)^{+}\big(I - \exp(-tX^T X/n)\big)X^T y$$

  - Then, for any time $t \geq 0$ (note the correspondence with λ),

$\mathrm{Bias}^2(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) \leq \mathrm{Bias}^2(\hat{\beta}^{\mathrm{ridge}}(1/t); \beta_0)$ and

$\mathrm{Var}(\hat{\beta}^{\mathrm{gf}}(t)) \leq 1.6862 \cdot \mathrm{Var}(\hat{\beta}^{\mathrm{ridge}}(1/t))$, so that

$\mathrm{Risk}(\hat{\beta}^{\mathrm{gf}}(t); \beta_0) \leq 1.6862 \cdot \mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(1/t); \beta_0)$

Excess risk bound (over ridge)

- Thm.: for any time $t > 0$ (provided the step size is small enough),

$$\begin{aligned}
&\mathrm{Risk}(\hat{\beta}^{\mathrm{sgf}}(t);\beta_0) - \mathrm{Risk}(\hat{\beta}^{\mathrm{ridge}}(1/t);\beta_0) \\
&\quad \le 0.6862 \cdot \mathrm{Var}_\eta(\hat{\beta}^{\mathrm{ridge}}(1/t)) && \text{(scaled ridge variance)} \\
&\qquad + \epsilon \cdot \frac{n}{m} \sum_{i=1}^{p} \mathbb{E}_\eta \left[\frac{\exp(\delta_y) s_i}{s_i - \alpha/2} \big(\exp(-\alpha t) - \exp(-2ts_i) \big) \right] \\
&&& \text{(“price of stochasticity”)} \\
&\qquad + \epsilon \cdot \frac{n}{m} \sum_{i=1}^{p} \mathbb{E}_\eta \Big[\gamma_y \big(1 - \exp(-2ts_i) \big) \Big] && \text{(limiting opt. error)}
\end{aligned}$$

- ϵ, m denote the step size and mini-batch size, respectively
- s_i denote the eigenvalues of the sample covariance matrix
- $\alpha, \gamma_y, \delta_y$ depend on $n, p, m, \epsilon, s_i, y$, but not t (see paper for details)

Implications/observations

- The second and third (variance) terms ...
  - Roughly scale with ϵ/m (Goyal et al., 2017; Smith et al., 2017; You et al., 2017; Shallue et al., 2019); this is different from gradient flow
  - Depend on the signal-to-noise ratio; this is different from gradient flow (and linear smoothers in general, because stochastic gradient flow/descent are actually *randomized* linear smoothers)
  - The second term decreases with time, just as a bias would; this is different from gradient flow (see lemma in the paper)

Implications/observations

- The second and third (variance) terms ...
  - Roughly scale with ϵ/m (Goyal et al., 2017; Smith et al., 2017; You et al., 2017; Shallue et al., 2019); this is different from gradient flow
  - Depend on the signal-to-noise ratio; this is different from gradient flow (and linear smoothers in general, because stochastic gradient flow/descent are actually *randomized* linear smoothers)
  - The second term decreases with time, just as a bias would; this is different from gradient flow (see lemma in the paper)

- Proof builds on the grad flow result, and uses the special covariance structure of the diffusion coefficient $Q_\epsilon(\beta(t))$ for least squares
  - Result(s) hold across the entire optimization path
  - No strong conditions placed on the data matrix X
  - Also, have the following lower bound under oracle tuning

$$\inf_{\lambda \geq 0} \operatorname{Risk}(\hat{\beta}^{\text{ridge}}(\lambda); \beta_0) \leq \inf_{t \geq 0} \operatorname{Risk}(\hat{\beta}^{\text{sgf}}(t); \beta_0)$$

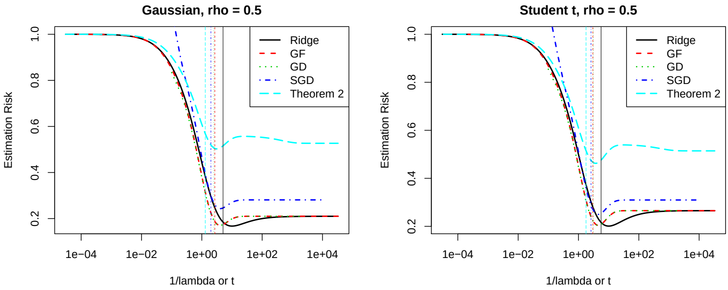
  - Similar result holds for the coefficient error (see theorem in paper)

$$\mathbb{E}_{\eta, Z} \|\hat{\beta}^{\text{sgf}}(t) - \hat{\beta}^{\text{ridge}}(1/t)\|_2^2$$

Outline

Synthetic data

- Below, we show $n = 100, p = 10, m = 2$
  - The bound (Theorem 2) tracks ridge's (and SGD's) risk(s) closely
  - The bound / SGD achieve risk comparable to grad flow in less time
  - See paper for other settings (e.g., high dimensions), coefficient error



Outline

Conclusion

- Gave theoretical and empirical evidence showing stochastic gradient flow is closely related to ℓ_2 regularization
- Interesting directions for future work
  - Showing that stochastic gradient flow and SGD are, in fact, close
  - Making the computational-statistical trade-off precise
  - General convex losses
  - Adaptive stochastic gradient methods

Thanks for listening!