ESCAPING SADDLE POINTS IN NON-CONVEX OPTIMIZATION

Γεώργιος Γκοτζιάς

GRADIENT DESCENT

$$X_{t+1} = x_t - \eta \, \nabla f(x_t)$$

- Αν η f είναι κυρτή, τότε συγκλίνει σε περιοχή γύρω από σημείο με $\nabla f(x_t) = 0$ (stationary point).
- Σε κάθε κυρτή συνάρτηση, αυτό το σημείο είναι το (ολικό) ελάχιστο.
- Αν η συνάρτηση δεν είναι κυρτή, Gradient Descent συγκλίνει σε stationary point.
- Όμως, stationary point μπορεί να είναι saddle point ή τοπικό μέγιστο ή τοπικό ελάχιστο.

GRADIENT DESCENT

- Πιθανές αλλαγές στον προηγούμενο αλγόριθμο, για να αποφεύγονται τα saddle points.

➢ Intermittent Perturbations: Πρόσθεση τυχαίας διαταραχής, υπό κάποια συνθήκη, στον GD.

➢ Random Initialization: Αρχικοποιούμε τον αλγόριθμο GD τυχαία.

- Και οι 2 αλλαγές διασφαλίζουν ότι αποφεύγονται τα saddle points.
- Όμως, με Random Initialization δεν διασφαλίζεται ότι θα γίνει αποδοτικά.

PERTURBED GRADIENT DESCENT

1. **For** t = 1, 2, … **do**
2. $x_t \leftarrow x_{t-1} - \eta \nabla f(x_{t-1})$
3. **If** perturbation condition holds **then**
4. $x_t \leftarrow x_t + \xi_t$

- Όπου ξ_t επιλέγεται ομοιόμορφα από μπάλα με κέντρο στο μηδέν και μικρή ακτίνα.
- Η συνθήκη υποθέτει μικρή τιμή για το $\nabla f(x_t)$

SADDLE POINTS

- Με τον όρο Saddle points περιλαμβάνουμε και τα τοπικά μέγιστα.
- Για κάθε Saddle Point, το Gradient είναι ίσο με μηδέν και η μικρότερη ιδιοτιμή του Hessian Matrix είναι μη θετική.
- Υπάρχει τουλάχιστον μία κατεύθυνση που είναι τοπικά μέγιστα.
- Αν η μικρότερη ιδιοτιμή του Hessian Matrix είναι μηδέν, δεν μπορούμε να συμπεράνουμε άμεσα αν είναι saddle point ή τοπικό ελάχιστο.
- Αν η μικρότερη τιμή του Hessian Matrix είναι αρνητική, τότε έχουμε strict saddle point.

ΟΡΙΣΜΟΙ

- Η συνάρτηση f είναι ℓ-gradient Lipschitz, αν
  - Για κάθε x_1, x_2, $|\nabla f(x_1) - \nabla f(x_2)| \leq \ell |x_1 - x_2|$
- Η συνάρτηση f είναι ρ-Hessian Lipschitz, αν
  - Για κάθε x_1, x_2, $||\nabla^2 f(x_1) - \nabla^2 f(x_2)|| \leq \rho |x_1 - x_2|$
- Το σημείο x είναι second-order stationary point, αν
  - $\nabla f(x) = 0$ και $\lambda_{min}(\nabla^2 f(x)) \geq 0$
- Το σημείο x είναι ε-second-order stationary point, αν
  - $|\nabla f(x)| \leq \varepsilon$ και $\lambda_{min}(\nabla^2 f(x)) \geq -\sqrt{(\rho\varepsilon)}$

MAIN THEOREM

- Αν η συνάρτηση f είναι ℓ-gradient Lipschitz και ρ-Hessian Lipschitz, τότε ο αλγόριθμος Perturbed Gradient Descent με η = O(1/ℓ), βρίσκει ε-second-order stationary point με μεγάλη πιθανότητα μετά από $O(\ell\,(f(x_0)-f^*)/\varepsilon^2)$ επαναλήψεις, αν αγνοήσουμε πολυώνυμα λογαριθμικών όρων.
- Αν η συνάρτηση, έχει μόνο strict saddle points, τότε second-order stationary points είναι τοπικά ελάχιστα.
- Το πλήθος επαναλήψεων είναι (ασυμπτωτικά) το ίδιο με αυτό για τον Gradient Descent για κυρτές συναρτήσεις (αγνοώντας λογαριθμικούς παράγοντες).

PERTURBED GRADIENT DESCENT

1. $t_{noise} \leftarrow - t_{thres} - 1$
2. **For** $t = 0, 1, 2, \ldots$ **do**
3. **If** $||\nabla f(x_t)|| \leq g_{thres}$ and $t - t_{noise} > t_{thres}$ **then**
4. $x'_t \leftarrow x_t$, $t_{noise} \leftarrow t$
5. $x_t \leftarrow x'_t + \xi_t$
6. **If** $t - t_{noise} = t_{thres}$ and $f(x_t) - f(x'_{tnoise}) > -f_{thres}$ **then**
7. **Return** x'_{tnoise}
8. $x_t \leftarrow x_{t-1} - \eta \nabla f(x_{t-1})$

PROOF OF MAIN THEOREM

- Αν είμαστε σε σημείο x_t τέτοιο, ώστε $||\nabla f(x_t)|| > g_{thres}$, για $\eta < 1/\ell$, ισχύει

$$f(x_{t+1}) \leq f(x_t) - (\eta/2)\ ||\nabla f(x_t)||^2$$

- Αν είμαστε σε σημείο x_t τέτοιο, ώστε $||\nabla f(x_t)|| \leq g_{thres}$ και $\lambda min(\nabla^2 f(x)) < - \sqrt{(\rho\varepsilon)}$ τότε θα έχουμε μία προσθήκη διαταραχής και ακολουθούν t_{thres} βήματα Gradient Descent, οπότε με μεγάλη πιθανότητα ισχύει ότι

$$f(x_{t+tthres}) - f(x_t) \leq - f_{thres}$$

COROLLARY

- Αν η συνάρτηση f είναι (θ, γ, ζ)-strict saddle, δηλαδή για κάθε x ισχύει ένα από τα ακόλουθα:

➢ $||\nabla f(x)|| \geq \theta$

➢ $\lambda_{min}(\nabla^2 f(x)) \leq -\gamma$

➢ X απέχει κατά ζ από κάποιο τοπικό ελάχιστο

- Αν η συνάρτηση f είναι ℓ-gradient Lipschitz, ρ-Hessian Lipschitz και (θ, γ, ζ)-stict saddle, τότε ο αλγόριθμος Perturbed Gradient Descent για $\varepsilon = \min(\theta, \gamma^2 / \rho)$, βρίσκει σημείο που απέχει κατά ζ από κάποιο τοπικό ελάχιστο με μεγάλη πιθανότητα μετά από $O(\ell\,(f(x_0)-f^*)/\varepsilon^2)$ επαναλήψεις, αν αγνοήσουμε πολυώνυμα λογαριθμικών όρων.

PROOF OF COROLLARY

- Από προηγούμενο θεώρημα, καταλήγουμε σε ε-second-order stationary point, οπότε $|\nabla f(x)| \le \varepsilon$ και $\lambda_{min}(\nabla^2 f(x)) \ge -\sqrt{(\rho\varepsilon)}$
- Όμως, $\varepsilon = \min(\theta, \gamma^2 / \rho)$, δηλαδή $\varepsilon \le \theta$ και $\varepsilon \le \gamma^2 / \rho$.
- Άρα, $|\nabla f(x)| \le \theta$ και $\lambda_{min}(\nabla^2 f(x)) \ge -\gamma$
- Και επειδή η f είναι (θ, γ, ζ)-strict saddle ικανοποιείται μόνο η τρίτη συνθήκη για το σημείο x.

ΣΥΜΠΕΡΑΣΜΑΤΑ- ΠΑΡΑΤΗΡΗΣΕΙΣ

- Ο αλγόριθμος Gradient Descent δουλεύει για μη κυρτές συναρτήσεις, εφόσον προστεθούν διαταραχές.
- Το πλήθος των επαναλήψεων είναι ανάλογο με ένα πολυώνυμο του λογαρίθμου του πλήθους των διαστάσεων.
- Ο αλγόριθμος Perturbed Gradient Descent βρίσκει σημείο το οποίο ικανοποιεί μία συνθήκη για το Hessian Matrix του, χωρίς να υπολογίζει ρητά τον πίνακα.