Résumé de TP : Optimisation Stochastique et Estimation de l'Espérance

Enzo Munoz, Maxime Guilbaud, Hugo Munier

1 Objectifs et Fondements Théoriques

L'objectif central de ce TP est de démontrer que l'espérance d'un vecteur aléatoire $X \in \mathbb{R}^d$ peut être identifiée comme la solution d'un problème d'optimisation, puis de comparer empiriquement la convergence de différents algorithmes de descente de gradient.

1.1 Démonstration de la fonction objectif

On considère la fonction objectif f définie par :

$$f(x_1, \dots, x_d) = \mathbb{E}\big[(X_1 - x_1)^2 + \cdots + (X_d - x_d)^2\big]$$

Le calcul du gradient donne :

$$\nabla f(x) = -2\big(\mathbb{E}[X_1] - x_1, \dots, \mathbb{E}[X_d] - x_d\big) = 2(x - \mathbb{E}[X])$$

Le point critique $\nabla f(x) = 0$ implique directement $x = \mathbb{E}[X]$. La matrice Hessienne étant $\nabla^2 f(x) = 2I_d$ (définie positive), la fonction est strictement convexe et admet $\mathbb{E}[X]$ comme unique minimum global.

2 Méthodologie et Protocoles de Simulation

L'étude compare deux algorithmes principaux pour minimiser f à partir de gradients bruités :

— **SGD (Stochastic Gradient Descent) :** Mise à jour classique avec un pas γ.

— **Adam :** Méthode adaptative utilisant les moments d'ordre 1 (β_1) et 2 (β_2) pour ajuster le pas par composante.

2.1 Lois de probabilité étudiées

Les modèles ont été testés sous trois configurations :

1. **Loi Normale :** Bruit uniforme sur toutes les dimensions.
2. **Loi de Student :** Distribution à queues lourdes pour évaluer la robustesse face aux valeurs extrêmes (bruit intense).

3 Analyse des Résultats et Observations

3.1 Sensibilité du paramètre de momentum β_1 (Adam)

L'étude détaillée de β_1 révèle un effet non monotone :

— **Valeurs faibles :** L'algorithme devient trop sensible au bruit instantané.

— **Valeurs élevées :** Une inertie excessive s'installe, empêchant une convergence précise vers la moyenne.

Il existe donc un compromis optimal pour β_1 permettant de stabiliser la trajectoire sans sacrifier la précision.

3.2 Décomposition Biais–Variance

L'analyse montre que l'erreur quadratique est largement **dominée par le biais** dans le cadre de ce TP. La variance reste relativement faible pour toutes les méthodes. Adam agit efficacement en réduisant l'impact du bruit sur la trajectoire, et parvient à supprimer totalement le biais pour un nombre fini d'itérations.

4 Conclusion

Dans ce TP, nous avons étudié différentes méthodes de descente de gradient stochastique pour estimer la moyenne d'une variable aléatoire, d'abord dans le cas d'une loi normale puis dans celui d'une loi de Student. Nous avons vu que le choix du pas joue un rôle central, en particulier pour l'algorithme de Robbins–Monro, où un pas mal choisi peut soit ralentir fortement la convergence, soit augmenter fortement la variance. Ce phénomène est encore plus marqué dans le cas de la loi de Student, en raison des queues lourdes qui génèrent des perturbations importantes dans le gradient.

L'algorithme AdaGrad permet de stabiliser la descente en adaptant automatiquement le pas en fonction des gradients observés. Cette adaptation améliore la robustesse par rapport à Robbins–Monro, mais dans le cas de la loi de Student, la présence fréquente de gradients de grande amplitude conduit à une diminution rapide du pas effectif. L'algorithme devient alors trop conservateur et n'atteint pas complètement la solution si le pas initial est trop petit, ce qui se traduit par la présence d'un biais résiduel.

L'algorithme Adam combine une adaptation du pas et un terme de momentum. Nos résultats montrent qu'il est globalement plus robuste que Robbins–Monro et AdaGrad, notamment en présence d'un bruit à queues lourdes. Toutefois, l'étude détaillée du paramètre de momentum β_1 met en évidence que son effet n'est pas monotone. Des valeurs trop faibles rendent l'algorithme trop sensible au bruit, tandis que des valeurs trop élevées introduisent une inertie excessive qui peut empêcher certaines composantes de converger précisément vers la moyenne. On observe ainsi l'existence d'un compromis, avec des valeurs intermédiaires de β_1 qui donnent les meilleures performances.

Enfin, l'analyse biais–variance montre que, dans ce cadre, l'erreur est largement dominée par le biais, la variance restant relativement faible pour l'ensemble des méthodes étudiées. Adam agit donc principalement en réduisant l'impact du bruit sur la trajectoire, mais sans supprimer totalement le biais pour un nombre fini d'itérations. Ce TP met ainsi en évidence à la fois l'intérêt des méthodes adaptatives et leurs limites, et montre que, même pour un algorithme robuste comme Adam, le choix des hyperparamètres reste important, en particulier en présence de distributions à queues lourdes.