Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Segundo Semestre 2024
Fundamentos Matemáticos para Ciencia de Datos IMT3120

Tarea 4

Instrucciones. Las preguntas 1 y 2 deben entregarse el 12 de Noviembre. La pregunta 3 debe entregarse el 24 de Noviembre. Para la parte computacional **no puede utilizar la librería scikit-learn** de python o alguna equivalente. Puede usar en cambio la librería numpy.

Pregunta 1 (40 %)**.** Consideramos un problema de optimización $\min\{f(w) : w \in \mathcal{W}\}$, con $f : \mathbb{R}^d \mapsto \mathbb{R}$ convexa y $\mathcal{W} \subseteq \mathbb{R}^d$ convexo y cerrado. Consideramos $G : \mathcal{W} \times \Xi \mapsto \mathbb{R}^d$ un oráculo de subgradiente estocástico para f, y usaremos ξ para denotar una semilla aleatoria para el oráculo.

En muchos problemas de aprendizaje no conocemos bien la geometría del problema. Por ejemplo, un cambio de base o re-escalamiento de las coordenadas puede permitir una optimización mucho más efectiva. El objetivo de esta pregunta es diseñar un algoritmo que re-escale las coordenadas de manera automática. Para esto, se propone el siguiente algoritmo:

Algoritmo 1 Método de Subgradiente Estocástico Adaptativo

1: **Input**: Modelo inicial $w^0 \in \mathbb{R}^d$, paso $\eta > 0$, parámetro de escalamiento $\delta > 0$, límite de iteraciones $T \in \mathbb{N}$, precisión $\epsilon > 0$
2: $R_{-1} = 0 \in \mathbb{R}^{d\times d}$
3: **while** $t \leq T$ **do**
4: $\quad g_t = G(w^t, \xi^{t+1})$
5: $\quad R_t = R_{t-1} + g_t g_t^\top$
6: $\quad H_t = \mathrm{Diag}(R_t)^{1/2}$
7: $\quad w^{t+1} = \arg\min\{\|w - [w^t - \eta H_t^{-1} g_t]\|_{H_t}^2 : w \in \mathcal{W}\}$
8: **end while**
9: **Output**: $\bar{w}^T = \frac{1}{T}\sum_{t=0}^{T-1} w^t$

Para este algoritmo (y su análisis), introducimos el producto interno y norma

$$\langle x, y\rangle_H = x^\top H y, \qquad \|x\|_H = \sqrt{\langle x, x\rangle}.$$

Recordemos además que $\mathcal{F}_t$ es el conjunto de eventos determinados por las variables aleatorias $\xi_1, \ldots, \xi_t$.

(a) El paso 7 de arriba corresponde a una proyección en la norma H_t (que denotamos $\Pi_{\mathcal{W}}^H(\cdot)$). Pruebe que $\Pi_{\mathcal{W}}^H(\cdot)$ es no-expansivo para $\|\cdot\|_H$; es decir:

$$\|\Pi_{\mathcal{W}}^H(a) - \Pi_{\mathcal{W}}^H(b)\|_H \leq \|a - b\|_H.$$

(b) Pruebe que si $w^* \in \arg\min\{f(w) : w \in \mathcal{W}\}$, entonces

$$\|w^{t+1} - w^*\|_{H_t}^2 \leq \|w^t - w^*\|_{H_t}^2 + \eta^2 \|g_t\|_{H_t^{-1}}^2 - 2\eta\langle g_t, w^t - w^*\rangle.$$

(c) Pruebe que

$$\mathbb{E}[\langle g_t, w^* - w^t\rangle | \mathcal{F}_t] \leq f(w^*) - f(w^t).$$

y concluya que

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*)$$
$$\leq \mathbb{E}\Big[\frac{1}{2\eta T}\Big(\|w^0 - w^*\|_{H_0}^2 + \sum_{t=1}^{T-1}\big(\|w^t - w^*\|_{H_t}^2 - \|w^t - w^*\|_{H_{t-1}}^2\big)\Big) + \frac{\eta}{2T}\sum_{t=0}^{T-1}\|g_t\|_{H_t^{-1}}^2\Big].$$

(d) Pruebe que si $D_\infty = \sup_{w\in\mathcal{W}} \|w - w^*\|_\infty$, entonces

$$\sum_{t=1}^{T-1}\big(\|w^t - w^*\|_{H_t}^2 - \|w^t - w^*\|_{H_{t-1}}^2\big) \leq D_\infty^2 \mathrm{tr}(H_{T-1} - H_0),$$

y que

$$\sum_{t=0}^{T-1}\|g_t\|_{H_t^{-1}}^2 \leq \mathrm{tr}(H_{T-1}).$$

Indicación. Para la segunda cota, pruebe que si h es una función monótona no-creciente, $(a_t)_{t=0,\ldots,T-1}$ sucesión de números no-negativos, y $A_t = \sum_{s=0}^t a_s$, entonces

$$\sum_{t=0}^{T-1} a_t h\Big(\sum_{s=0}^t a_s\Big) \leq \int_0^{A_{T-1}} h(x)\,dx.$$

(e) Concluya que para $\eta = D_\infty$, tenemos

$$\mathbb{E}[f(\bar{w}^T)] - f(w^*) \leq \mathbb{E}\Big[\frac{D_\infty \mathrm{tr}(H_{T-1})}{T}\Big].$$

Suponga ahora que el oráculo posee cota de segundo momento $\nu^2 \geq 0$. Compare la cota de arriba con aquella probada en clases para el método de subgradiente estocástico con paso constante.

Pregunta 2 (20 %)**.** Suponga ahora adicionalmente que el oráculo de gradiente estocástico está asociado a una función: es decir, existe $f(\cdot,\cdot) : \Omega \times \Xi \mapsto \mathbb{R}$ tal que

$$\mathbb{E}_\xi[f(w,\xi)] = f(w), \qquad G(w,\xi) = \nabla f(w,\xi).$$

y suponemos adicionalmente que

- $f(w,\xi) \in [-M, +M]$, para todo $w \in \Omega$, ξ.
- $\sup_{w,\xi} \|G(w,\xi)\|_2 \leq \hat{L}$

Pruebe la mejor cota de alta probabilidad para el Algoritmo 1 que pueda, bajo los supuestos anteriores.

Tarea Parte Computacional (40 %)

Implemente los siguientes métodos:

1. Método de Subgradiente Estocástico Adaptativo.
2. Método de Subgradiente Estocástico de Langevin con paso fijo. Para este método, puede modificar el código hecho en la tarea anterior (si utilizó la librería scikit-learn, deberá hacer las modificaciones necesarias para no depender de esta librería).

Para ambos algoritmos, la clase de modelos son redes neuronales con una capa oculta, activación ReLU y pérdida logística (al igual que en la tarea anterior).

Para evaluar sus modelos, utilice el conjunto de datos RCV1-v2 (Lewis, Yang, Rose, Li:JMLR 2004)

`https://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf`

Este conjunto de datos posee aproximadamente 800.000 artículos de la empresa periodística Reuters, y posee etiquetas correspondientes a 4 categorías de alto nivel: Economía, Comercio, Medicina, and Gobierno (ECAT, CCAT, MCAT, GCAT). Tendremos en cuenta las siguientes consideraciones:

- Dado que en este conjunto de datos el conjunto de entrenamieno (23.149 documentos) es más pequeño que el de validación (781.265), revierta los roles de ambos conjuntos de datos.
- Extraiga 20.000 documentos al azar del conjunto de 781.265 documentos para construir un conjunto de validación. Con este conjunto se hará la selección de hiperparámetros.
- El conjunto de testeo será el de 23.149 documentos.

Pregunta 3. (a) Utilice el conjunto de entrenamiento (781.265-20.000) para todos los conjuntos de hiperparámetros ($\eta \in \{10^{-4}, 10^{-3}, 10^{-2}\}$, tamaño de batch, ancho de la capa oculta, y parámetro de regularización), escoja el modelo que obtenga el mejor desempeño en el conjunto de validación.

(b) Grafique, para dichos hiperparámetros, la curva de entrenamiento y testeo de su algoritmo, y compare ambos algoritmos.

(c) Para los hiperparámetros escogidos, haga una evaluación empírica de la estabilidad algorítmica de ambos métodos. Para esto:

- Escoja un dato del conjunto dez entrenamiento (por ejemplo, el primero).
- Reemplace sus atributos y etiqueta por 0.
- Evalúe la distancia $\{\|w^t - v^t\|_2 : t \in [T]\}$, cuando realiza el entrenamiento con el conjunto de datos original, y el conjunto modificado en el dato escogido.

De acuerdo al experimento: parecen los métodos algorítmicamente estables? Es la distancia de los modelos monótona en el número de iteraciones? Hay una comparación clara en la estabilidad de ambos métodos? Note que en el caso de redes neuronales, las pérdidas son no-convexas y no-suaves, por lo que ninguna de las hipótesis vistas en clase se cumple.