NIPS, December 4-9, 2017, Long Beach

Safe Adaptive Importance Sampling

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
Max Planck Institute for Intelligent Systems, Tübingen Campus
sebastian.stich@epfl.ch, anant.raj@tuebingen.mpg.de, martin.jaggi@epfl.ch

Summary

Optimal adaptive sampling (full information)
The performance of stochastic optimization algorithms like Stochastic Gradient Descent (SGD) and Coordinate Descent (CD) crucially depends on the sampling distribution. The progress is maximized for

full information optimal adaptive sampling $\mathbf{p}^{\text{opt}}$,

but this distribution requires knowledge of full gradient information and is therefore unamenable in practice.

Safe bounds: a relaxation

CD setting:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \quad \Rightarrow \quad [\boldsymbol{\ell}]_i \leq |\nabla_i f(\mathbf{x})| \leq [\mathbf{u}]_i$$

SGD setting:

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x}) \quad \Rightarrow \quad [\boldsymbol{\ell}]_i \leq \|\nabla f_i(\mathbf{x})\| \leq [\mathbf{u}]_i$$

Optimal adaptive sampling (limited information)
The optimal adaptive sampling with respect to the bounds $\boldsymbol{\ell} \leq \mathbf{u}$ is the solution of an **optimization problem** and can efficiently be computed. We propose to use

limited information optimal adaptive sampling $\tilde{\mathbf{p}}^{\boldsymbol{\ell},\mathbf{u}}$.

For any bounds $\boldsymbol{\ell} \leq \mathbf{u}$, the proposed sampling $\tilde{\mathbf{p}}^{\boldsymbol{\ell},\mathbf{u}}$ is provably better than importance sampling.

Example: Coordinate Descent (also applicable to SGD)

Alg. 1: Optimal sampling (too expensive)
- **Compute:** $\nabla f(\mathbf{x}_k)$
- $\boldsymbol{\ell}_k = \mathbf{u}_k = |\nabla f(\mathbf{x}_k)|$
- **Iteration complexity:** $T(\nabla f(\mathbf{x}_k))$

Alg. 2: Proposed sampling
- **Compute:** $\nabla_i f(\mathbf{x}_k)$
- Update $\boldsymbol{\ell}_k \leq |\nabla f(\mathbf{x}_k)| \leq \mathbf{u}_k$
- **Iteration complexity:** $T(\nabla_i f(\mathbf{x}_k)) + O(n \log n)$

Alg. 3: Importance sampling (slower convergence)
- **Compute:** $\nabla_i f(\mathbf{x}_k)$
- $\boldsymbol{\ell}_k = \mathbf{0}, \mathbf{u}_k = \boldsymbol{\infty}$
- **Iteration complexity:** $T(\nabla_i f(\mathbf{x}_k))$

- Compute $(v_k, \mathbf{p}_k) = \textbf{OptimalSampling}(\boldsymbol{\ell}_k, \mathbf{u}_k)$
- select index $i_k \sim \mathbf{p}_k$ and update

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{v_k [\mathbf{p}_k]_{i_k}} \nabla_{i_k} f(\mathbf{x}_k)$$

The stepsize $\frac{1}{v_k}$ and the sampling $\mathbf{p}_k$ maximize the expected one step progress (in the worst case):

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq \underbrace{\frac{\langle \nabla f(\mathbf{x}_k), \nabla_{i_k} f(\mathbf{x}_k)\rangle}{v_k [\mathbf{p}_k]_{i_k}} - \frac{L_{i_k} \|\nabla_{i_k} f(\mathbf{x}_k)\|^2}{2 v_k^2 [\mathbf{p}_k]_{i_k}^2}}_{:=\tau(v_k, \mathbf{p}_k, \nabla f(\mathbf{x}_k))}$$

$$(v_k, \mathbf{p}_k) := \arg\max_{\substack{v\in\mathbb{R} \\ \mathbf{p}\in\Delta^n}} \min_{\boldsymbol{\ell}_k \leq \mathbf{c} \leq \mathbf{u}_k} \mathbb{E}_{i_k\sim\mathbf{p}} [\tau(v, \mathbf{p}, \mathbf{c})]$$

Alg. 1: Optimal sampling
- $\mathbf{p}_k^{\text{opt}} := \frac{|\sqrt{\mathbf{L}}\nabla f(\mathbf{x}_k)|}{\|\sqrt{\mathbf{L}}\nabla f(\mathbf{x}_k)\|_1}$, $v_k^{\text{opt}} = \frac{\|\sqrt{\mathbf{L}}\nabla f(\mathbf{x}_k)\|_1^2}{\|\nabla f(\mathbf{x}_k)\|_2^2}$

Alg. 2: Proposed sampling
- **Theorem 3.4**: $\tilde{\mathbf{p}}_k$, $\tilde{v}_k$ can be computed in $O(n \log n)$
- **Theorem 3.2**: $v_k^{\text{opt}} \leq \tilde{v}_k \leq v_k^{\text{imp}}$ (the progress is always better than importance sampling)
- $\tilde{\mathbf{p}}_k$ is the *best* sampling (in worst case) for $\boldsymbol{\ell}_k, \mathbf{u}_k$

Alg. 3: Importance sampling
- $\mathbf{p}_k^{\text{imp}} := \frac{|\mathbf{L}|}{\text{Tr}[\mathbf{L}]}$, $v_k^{\text{imp}} = \text{Tr}[\mathbf{L}]$
- Note: for $L_i \equiv L$, this is just uniform sampling

Assumptions: $f\colon \mathbb{R}^n \to \mathbb{R}$ convex; coordinate-wise Lipschitz: $|\nabla_i f(\mathbf{x} + \gamma \mathbf{e}_i) - \nabla_i f(\mathbf{x})| \leq L_i |\gamma|$, $\forall \mathbf{x} \in \mathbb{R}^n, \gamma \in \mathbb{R}, i = 1\colon n$. Define: $\mathbf{L} = \text{diag}(L_1, \ldots, L_n)$.

Details

- Trivial values $[\boldsymbol{\ell}]_i = 0$ and $[\mathbf{u}]_i = \infty$ are admissible, but more accurate bounds give better speed-up.
- Updating the bounds can be delegated to a dedicated worker in a distributed setting.

Special cases:
- $\boldsymbol{\ell} = \mathbf{u}$: optimal sampling (full information)
- $\boldsymbol{\ell} = \mathbf{0}$, $\mathbf{u} = \boldsymbol{\infty}$: uniform sampling (no information)

Principal example: $f(\mathbf{x}) = \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|^2$.
- **CD setting:** $(A \in \mathbb{R}^{d\times n})$

$$\nabla_i f(\mathbf{x} + \gamma \mathbf{e}_{i_k}) - \nabla_i f(\mathbf{x}) = \gamma \langle \mathbf{a}_i, \mathbf{a}_{i_k}\rangle, \quad \forall i \neq i_k$$

- **SGD setting**: $(A \in \mathbb{R}^{n\times d})$, $f(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{a}_i^\top \mathbf{x})$

$$\nabla f_i(\mathbf{x} + \gamma \mathbf{e}_{i_k}) - \nabla f_i(\mathbf{x}_k) = \gamma \langle \mathbf{a}_i, \mathbf{a}_{i_k}\rangle \mathbf{a}_i, \quad \forall i \neq i_k$$

- $T(\nabla f(\mathbf{x})) = \Theta(dn)$, $T(\nabla_i f) = T(\nabla f_i) = \Theta(d)$
- Updating the bounds takes $\Theta(n)$ time.

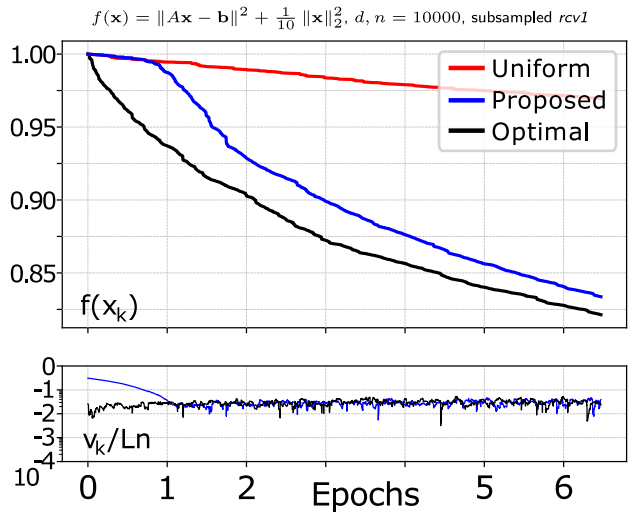
Upper/lower bounds alone are not enough:

Example 3.1: $\boldsymbol{\ell} = \begin{bmatrix}1\\2\end{bmatrix}$, $\mathbf{u} = \begin{bmatrix}2\\3\end{bmatrix}$, $\nabla f(\mathbf{x}_k) = \begin{bmatrix}2\\2\end{bmatrix}$ and $L_1 = L_2 = 1$. Then

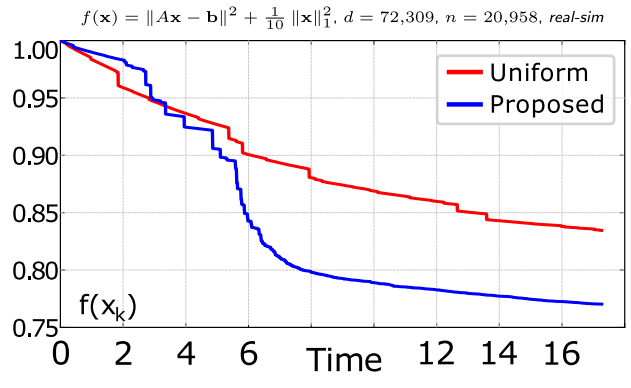
$$\mathbb{E}_{i_k\sim\text{uniform}} [f(\mathbf{x}_{k+1})] < \max_{\mathbf{p}=\left\{\frac{\mathbf{u}}{\|\mathbf{u}\|_1}, \frac{\boldsymbol{\ell}}{\|\boldsymbol{\ell}\|_1}\right\}} \mathbb{E}_{i_k\sim\mathbf{p}} [f(\mathbf{x}_{k+1})]$$

Experiments

Iterations: Coordinate Descent on *rcv1*



Clock time: Coordinate Descent on *real-sim*



Experiments with SGD can be found in the full paper.

Open Problems and Future Work

- The approach works very well for the CD setting, for SGD the advantage was less pronounced in the experiments, can this be explained by theory?
- Relax the strict conditions on $\boldsymbol{\ell}$ and $\mathbf{u}$
- Reduce complexity to $\tilde{O}(\min\{d, n\})$ (for both: updating the bounds $\boldsymbol{\ell}, \mathbf{u}$, and computing $\tilde{\mathbf{p}}^{\boldsymbol{\ell},\mathbf{u}}$).
- The approach might be transferable to other domains, for instance active learning.