An interior-point stochastic approximation method and an L1-regularized delta rule

Peter Carbonetto, Mark Schmidt and Nando de Freitas
University of British Columbia
A motivating example

• **Goal:** \( p(\text{spam} \mid \text{email-features}, \theta) \).
  (Cormack and Lynam, 2005)

• **Approach:** find model \( \theta \) that maximizes likelihood of training data.

• **But:** training data is observed over time (on-line learning).

• **Moreover:** we need to penalize complex models \( \theta \).
What is stochastic approximation, briefly

- **Original Problem:** (Spall, 2003; Kushner and Yin, 2003; Bottou, 1998)
  1. Minimize $f(x)$, or find $F(x) = \nabla f(x) = 0$.
  2. We only observe noisy, unbiased $g_k \approx F(x_k)$.

- **Robbins & Monro algorithm:**
  1. $x_{k+1} = x_k - a_k g_k$
  2. $\{a_k\}$ is a sequence of decreasing step sizes.

- **Problem in this talk:**
  
  minimize $f(x)$
  subject to $c(x) \leq 0$. 
Motivating example (continued)

• **nonsmooth, unconstrained** objective:

  minimize $-\log p(\text{spam} \mid \text{email-features}, \theta) + \lambda \| \theta \|_1$

• change $\theta$ to $x$ to obtain **smooth, constrained** objective:

  minimize $-\log p(\text{spam} \mid \text{email-features}, x) + \lambda \sum_i x_i$

  subject to $x \geq 0$.

\[
\|u\|_1 \equiv \sum_i |u_i|
\]
Projected gradient
(Bertsekas, 1999; Poljak, 1978)

✓ Has convergence guarantees.
✗ Not always efficient to compute projection.
✗ Big steps may be biased ⇒ slow progress.
The interior-point approach

Projected gradient

✓ Has convergence guarantees.

✗ Only feasible for simple constraints.

✗ Large steps may be biased.

Primal-dual Interior-point method

✓ Also has convergence guarantees.

✓ Works for many types of constraints.

✓ Steps are never biased.
The interior-point approach

- **Our problem:**
  
  minimize $f(x)$
  
  subject to $c(x) \leq 0$.

- **The barrier function:**
  
  (Fiacco and McCormick, 1968)
  
  $$f_\mu(x) \equiv f(x) - \mu \sum_{i=1}^{m} \log(-c_i(x))$$
The interior-point approach

- **Primal interior-point method:** solve sequence \( F_\mu (x) = \nabla f_\mu (x) = 0 \) for decreasing \( \mu > 0 \).

- **But:** we cannot assess convergence to each subproblem \( F_\mu (x) = 0 \)!

Adapted from Fiacco and McCormick (1968).
The “"primal-dual”” approach

(M. H. Wright, 1992; S. J. Wright, 1996; many others...)

- Recall problem: minimize $f(x)$ subject to $c(x) \leq 0$.
- Introduce Lagrange multiplier-like variables $z$.
- Use Robbins-Monro to solve moving target:

$$F_\mu(x, y) \equiv \begin{bmatrix} \nabla f(x) + \nabla c(x)^T z \\ c(x) \cdot z + \mu \end{bmatrix} = \begin{bmatrix} \text{gradient of Lagrangian} \\ \text{complementarity} \end{bmatrix} = 0.$$  

- Take steps:

$$x_{k+1} = x_k + \hat{a}_k \Delta x_k$$
$$z_{k+1} = z_k + \hat{a}_k \Delta z_k$$

Replace with “noisy” estimate

“Perturbed” KKT conditions

Primal-dual search direction
Why does this work?

1. Central path ⇒ numerically stable.

2. Primal-dual search direction keeps us on central path, even with noisy gradients.
A small experiment

• **Problem:** linear regression + L1 penalty (Lasso)

\[
\begin{align*}
\text{minimize} & \quad \|Ax - b\|^2 + \lambda \sum_i x_i \\
\text{subject to} & \quad x \geq 0.
\end{align*}
\]

• **Question:** how well does on-line estimate recover exact solution?

• Synthetic data.

• Repeat for \( k = 1 \) to 100, step sizes \( a_k = 1/(k_0 + k) \).
A small experiment

• Compared these methods:
  • Projected gradient
  • Primal-dual interior-point
  • Sub-gradient (Shalev-Shwartz et al, 2007; Hazan, 2007)
  • Smoothed approximation
  • Augmented Lagrangian (Wang and Spall, 2003)
Some empirical evidence

Sensitivity to step size sequence

(distance to exact solution vs. $k_0$)

(step sizes are $\alpha_k = 1/(k_0 + k)$.)
Some empirical evidence

Sensitivity to strength of L1 penalty

penalty strength

distance to exact solution

An interior-point stochastic approximation method and an L1-regularized delta rule
Shrinkage effect

- exact solution
- online 1 pass of data
- online 10 passes

regression coefficient vs. penalty strength
In summary

- Robbins-Munro solved $F(x) = 0$ with updates

$$x_{k+1} = x_k - a_k g_k.$$  

- We solve sequence $F_\mu (x,z) = 0$ with updates

$$x_{k+1} = x_k + \hat{a}_k \Delta x_k$$
$$z_{k+1} = z_k + \hat{a}_k \Delta z_k$$

where $(\Delta x, \Delta z)$ is the solution to the “perturbed” KKT conditions.
Thank you!