Almost surely constrained convex optimization

Ahmet Alacaoglu
ahmet.alacaoglu@epfl.ch

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

ICML 2019
Long Beach
[June 12, 2019]

Joint work with
Olivier Fercoq @ Telecom Paris-Tech, Ion Necoara @ UPB
Volkan Cevher @ LIONS
Almost surely constrained convex optimization:

\[
\begin{align*}
\min_{x \in \mathbb{R}^d} \{ P(x) := F(x) + h(x) \} \\
A(\xi)x \in b(\xi) \quad \xi\text{-almost surely,}
\end{align*}
\]

- \( F(x) = \mathbb{E}[f(x, \xi)] \), with convex and smooth \( f(\cdot, \xi) : \mathbb{R}^d \to \mathbb{R} \).
- \( h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\} \) is a nonsmooth, proximable convex function.
- \( A(\xi) \in \mathbb{R}^{m \times d} \) and \( b(\xi) \subseteq \mathbb{R}^m \) are random.

- Some applications: support vector machines, basis pursuit, portfolio optimization, semi-infinite programming...
Prior art

\[
\min_{x \in \mathbb{R}^d} \mathbb{E}[f(x, \xi)] : x \in \mathcal{B}(\xi) \cap \Omega
\]

• Idea: Use alternating projections for the constraints.

• Update:

\[
y^k = \text{prox}_{\mu_k f(\cdot, \xi)} \left( x^k \right) \]

\[
x^{k+1} = \text{proj}_{\mathcal{B}(\xi)}(y^k)
\]

• Drawbacks:
  o More restricted problem class.
  o Requires projectability of sets.

Primer on stochastic proximal gradient method (SPG)

\[
\min_{x \in \mathbb{R}^d} \{ P(x) := F(x) + h(x) \}
\]

- **SPG:**
  \[
x^{k+1} = \text{prox} \left( \frac{\alpha_0}{\sqrt{k}} h \left( x^k - \frac{\alpha_0}{\sqrt{k}} \nabla f(x^k, \xi) \right) \right).
\]

- **Convergence rate:**
  \[
P(x^k) - P(x^*) \leq O \left( \frac{\sigma^2 + L\|x^0 - x^*\|^2}{\sqrt{k}} \right).
\]
  - Standard assumption: Bounded variance:
    \[
    \mathbb{E} \| \nabla F(x) - \nabla f(x, \xi) \|^2 \leq \sigma^2 < \infty.
    \]
Primer on smoothing

• A smooth estimate of $g = \delta_b(\xi)$:

$$g_\beta(A(\xi)x, \xi) = \max_{y \in \mathbb{R}^m} \left\{ \langle A(\xi)x, y \rangle - g^*(y, \xi) - \frac{\beta}{2} \|y\|^2 \right\}.$$  

• $g_\beta$ is differentiable and $\nabla g_\beta$ is $\frac{1}{\beta}$-Lipschitz continuous.

$$g_\beta(A(\xi)x, \xi) = \frac{1}{2\beta} \text{dist}(A(\xi)x, b(\xi))^2$$

$$G_\beta(Ax) = \frac{1}{2\beta} \mathbb{E} \left[ \text{dist}(A(\xi)x, b(\xi))^2 \right],$$

where $\text{dist}(x, \mathcal{K}) = \inf_{y \in \mathcal{K}} \|x - y\|$.
Stochastic gradients of smoothed function

**Algorithmic Idea:** Apply SGD to

$$\min_{x \in \mathbb{R}^d} \left\{ P_\beta(x) := \mathbb{E} f(x, \xi) + h(x) + G_\beta(Ax) \right\},$$

with $\beta$ decreasing to 0.

- Recall:

  $$g_\beta(A(\xi)x, \xi) = \frac{1}{2\beta} \text{dist}(A(\xi)x, b(\xi))^2$$

  $$G_\beta(Ax) = \frac{1}{2\beta} \mathbb{E} \left[ \text{dist}(A(\xi)x, b(\xi))^2 \right].$$

• Taking stochastic gradients:

  $$\nabla_x g_\beta(A(\xi)x, \xi) = A(\xi)^\top \nabla_A(A(\xi)x) + \frac{1}{\beta} \text{dist}(A(\xi)x, b(\xi))^2.$$
Stochastic gradients of smoothed function

**Algorithmic Idea:** Apply SGD to

\[
\min_{x \in \mathbb{R}^d} \left\{ P_\beta(x) := \mathbb{E}f(x, \xi) + h(x) + G_\beta(Ax) \right\},
\]

with \( \beta \) decreasing to 0.

- Recall:

  \[
  g_\beta(A(\xi)x, \xi) = \frac{1}{2\beta} \operatorname{dist}(A(\xi)x, b(\xi))^2
  \]

  \[
  G_\beta(Ax) = \frac{1}{2\beta} \mathbb{E} \left[ \operatorname{dist}(A(\xi)x, b(\xi))^2 \right].
  \]

- Taking stochastic gradients:

  \[
  \nabla_x g_\beta(A(\xi)x, \xi) = A(\xi)^\top \nabla_{A(\xi)x} \frac{1}{2\beta} \operatorname{dist}(A(\xi)x, b(\xi))^2
  \]

  \[
  = \frac{1}{\beta} A(\xi)^\top (A(\xi)x - \operatorname{proj}_{b(\xi)}(A(\xi)x)).
  \]

- Only requires projections to \( b(\xi) \).

- Challenge: Standard variance bound does not hold as \( \beta \to 0 \).
SASC for general convex case

Input: \( x_0^0 \in \mathbb{R}^d \)

Parameters: \( \alpha_0 \leq \frac{3}{4L(\nabla F)} \), and \( \omega > 1 \)

\( m_0 \in \mathbb{N}_* \).

for \( s \in \mathbb{N} \) do

\( m_s = \lfloor m_0 \omega^s \rfloor \), and \( \alpha_s = \alpha_0 \omega^{-s/2} \).

\( \beta_s = 4\alpha_s \sup_{\xi} \|A(\xi)\|^2 \).

for \( k \in \{0, \ldots, m_s - 1\} \) do

Draw \( \xi = \xi_{k+1}^s \).

\( x_{k+1}^s = \text{prox}_{\alpha_s h} \left( x_k^s - \alpha_s \left[ \nabla f(x_k^s, \xi) + \frac{1}{\beta_s} A(\xi)^\top (A(\xi)x_k^s - \text{proj}_b(\xi)(A(\xi)x_k^s)) \right] \right) \)

end for

\( \bar{x}^s = \frac{1}{m_s} \sum_{k=1}^{m_s} x_k^s \)

\( x_{0+1}^s = x_{m_s}^s \).

end for

return \( \bar{x}^s \)
Lagrangian, primal-dual solutions

\[
\min_{x \in \mathbb{R}^d} \{ P(x) := F(x) + h(x) \} \\
A(\xi)x \in b(\xi) \quad \xi\text{-almost surely,}
\]

- Define the Lagrangian:

\[
\mathcal{L}(x, y) = P(x) + \int \langle A(\xi)x, y(\xi) \rangle - \sup_{b(\xi)} (y(\xi)) \mu(d\xi),
\]

where \( \sup_{\mathcal{K}} y = \sup_{x \in \mathcal{K}} \langle x, y \rangle \).

- \((x_\star, y_\star)\) is a saddle point of

\[
\min_{x \in \mathbb{R}^d} \max_{y \in \mathcal{Y}} \mathcal{L}(x, y).
\]
A key lemma

\[
\min_{x \in \mathbb{R}^d} \{ P(x) := F(x) + h(x) \}
\]

\[
A(\xi)x \in b(\xi) \quad \xi\text{-almost surely},
\]

- Define \( S_\beta(x) = P_\beta(x) - P(x_\star) = P(x) - P(x_\star) + \frac{1}{2\beta} \int \text{dist}(A(\xi)x, b(\xi))^2 \mu(d\xi). \)

Then, the following hold:

\[
S_\beta(x) \geq -\frac{\beta}{2} \|y_\star\|^2,
\]

\[
P(x) - P(x_\star) \geq -\frac{1}{4\beta} \int \text{dist}(A(\xi)x, b(\xi))^2 \mu(d\xi) - \beta \|y_\star\|^2,
\]

\[
P(x) - P(x_\star) \leq S_\beta(x),
\]

\[
\int \text{dist}(A(\xi)x, b(\xi))^2 \mu(d\xi) \leq 4\beta^2 \|y_\star\|^2 + 4\beta S_\beta(x). 
\]

If \( S_\beta \) and \( \beta \) are small, then objective residual and feasibility values are also small.
Main theorem

\[
\min_{x \in \mathbb{R}^d} \{P(x) := F(x) + h(x)\}
\]

\[A(\xi)x \in b(\xi) \quad \xi\text{-almost surely,}\]

- Denote by \(M_s = \sum_{i=0}^{s} m_i\) total number of iterations. Then, the iterates of SASC satisfy

\[
\mathbb{E}|P(\bar{x}^s) - P(x_*)| \leq O\left(\log_\omega (M_s/m_0) \frac{\sigma_f^2 + \|x_* - x_0\|^2 + \|y_*\|^2}{\sqrt{M_s}}\right),
\]

\[
\sqrt{\mathbb{E}[\text{dist}(A(\xi)\bar{x}^s, b(\xi))^2]} \leq O\left(\log_\omega (M_s/m_0) \frac{\sigma_f^2 + \|x_* - x_0\|^2 + \|y_*\|^2}{\sqrt{M_s}}\right).
\]

- This rate is optimal even without constraints up to a logarithmic factor.
Extensions: Restricted strongly convex

\[
\min_{x \in \mathbb{R}^d} \{ P(x) := F(x) + h(x) \}
\]

\[
A(\xi)x \in b(\xi) \quad \xi\text{-almost surely},
\]

- Denote by \( M_s = \sum_{i=0}^{s} m_i \) total number of iterations.
- If \( P(x) \) satisfies the quadratic growth condition:
  \[
P(x) - P(x_\star) \geq \frac{\mu}{2} \| x - x_\star \|^2,
  \]
  the iterates of SASC satisfy
  \[
  \mathbb{E} | P(\tilde{x}^s) - P(x_\star) | \leq \mathcal{O} \left( \log_\omega (M_s/m_0) \frac{\sigma_f^2 + \| x_\star - x_0 \|^2 + \| y_\star \|^2}{M_s} \right),
  \]
  \[
  \sqrt{\mathbb{E} \left[ \text{dist}(A(\xi)\tilde{x}^s, b(\xi))^2 \right]} \leq \mathcal{O} \left( \log_\omega (M_s/m_0) \frac{\sigma_f^2 + \| x_\star - x_0 \|^2 + \| y_\star \|^2}{M_s} \right).
  \]
- This rate is optimal even without constraints up to a logarithmic factor.
Numerical experiments: Basis pursuit

\[ \min_{x \in \mathbb{R}^d} \|x\|_1 \]

\[ \text{st: } a^\top x = b, \text{a.s.} \]

- Data generation:
  - \( \Sigma_{i,j} = \rho |i-j| \) with \( \rho = 0.9 \).
  - \( x^* \in \mathbb{R}^d \), \( d = 100 \) with 10 nonzero coefficients.
  - \( a_i \sim \mathcal{N}(0, \Sigma) \) independent random variables, which are then centered and normalized.
  - \( b_i = a_i^\top x^*, \ i \in [1, m] \) where \( m = 10^5 \).
- Because of the centering, there are multiple solutions to the infinite system \( a^\top x = b \) a.s.
Numerical experiments: Basis pursuit

- SGD does not converge to the sparse solution.
- SPP stagnates at the predefined accuracy, due to fixed step size.

Numerical experiments: SVM

Hard margin SVM:
\[
\min_{x \in \mathbb{R}^d} \frac{1}{2} \|x\|^2 : b_i \langle a_i, x \rangle \geq 1, \forall i.
\]

- SASC applies to hard margin SVM.

Soft margin SVM:
\[
\min_{x \in \mathbb{R}^d} \frac{1}{2} \|x\|^2 + C \sum_{i=1}^{n} \max \{0, 1 - b_i \langle a_i, x \rangle\},
\]

- Pegasos (primal subgradient method) applies to soft margin SVM.

Numerical experiments: SVM

Dataset 1: kdd2010: 19,264,997 training examples, 748,401 testing examples, 1,163,024 features
Dataset 2: news20: 17,996 training examples, 2,000 testing examples, 1,355.191 features
Dataset 3: rcv1: 20,424 training examples, 677,399 testing examples, 47,236 features

- Accuracy of Pegasos depends on the regularization parameter.
- SASC is comparable to Pegasos with the best regularization parameter.

Conclusions

- SGD-type method for stochastic optimization with infinitely many linear inclusion constraints.
- Optimal convergence rates up to a logarithmic factor.
- Extensions for solving

\[
\min_{x \in \mathbb{R}^d} \mathbb{E}[f(x, \xi) + g_1(A_1(\xi)x, \xi)] + h(x), \\
A_2(\xi)x \in b(\xi), \xi\text{-almost surely},
\]

with nonsmooth and Lipschitz continuous \( g_1 \).
- State-of-the-art practical performance.

To learn more: ahmet.alacaoglu@epfl.ch

- Poster @ Pacific Ballroom #101